OVERALL

SPECIFIC AIMS

The overall goal of the Drug Design Data Resource (D3R) is to create technologies that will enable dramatic advances in two core purposes of computer-aided drug design (CADD): ligand-protein pose prediction and affinity prediction or ranking. Success in this endeavor will lower the cost and accelerate the discovery of new medications across a range of therapeutic areas.

Our approach is based on a recognition that the field currently lacks effective methods to test the accuracy of CADD methods, and that such methods are needed to advance the state of the art. A major challenge is the shortage of unpublished data that can be used to carry out blinded -- and hence objective, large-scale -- and hence statistically significant -- tests. We propose to overcome this challenge by tapping into two large, existing flows of data in a manner that will generate effectively blinded prediction challenges two orders of magnitude larger than are possible today. The first is the flow of new PDB entries, many of which are cocrystal structures of proteins with drug-like ligands; these enable the pose-prediction challenge. The second is the flow of newly curated protein-ligand binding data generated by the BindingDB project; these will enable the affinity-prediction challenge. For pose-predictions, a blinded challenge is enabled by using the PDB's advance notice of forthcoming structures. For affinity predictions, a blinded challenge is enabled by ensuring that the predictions are made by automated workflows that are isolated from the internet before the affinity data become available.

Furthermore, CADD methods are increasingly complicated and thus, we need a way to dramatically increase the scale and throughput, as well as the rigor and reproducibility, of the methods employed. We see the development of automated workflows, encapsulating the entire end-to-end CADD experiment, as a key technology to embrace and enable. Thus, we propose to create software and workflow frameworks that will coordinate these continuous, blinded prediction challenges and automatically archive the results, and we will also disseminate and evaluate the results. This work will be done in concert with two classes of Driving Biomedical Project partners. The first class comprises CADD method developers who will use the new frameworks to test and improve their methods. The second class comprises researchers who will apply CADD methods that have been found to perform particularly well in their drug design projects. The latter class includes an innovative plan to crowd-source ligand design, much as FoldIt crowd-sourced protein design. We will engage also extensively with the research community to collect input and maximize the impact of this project.

We will achieve our goals through the following tightly integrated Specific Aims:

Aim 1 Develop, operate, learn from, and disseminate methods and results of high-throughput, continuous, blinded pose-prediction and affinity prediction challenges. Create workflow frameworks and model workflows, and work with DBPs to develop workflows. Pose throughput expected >1000/year; affinity throughput >10,000/year. Archive workflows, affinity data, and workflow predictions and evaluations for dissemination, reuse, and study analysis. (**TRDPs 1,2 and 3.**)

Aim 2 Support and collaborate with DBP partners utilizing the technologies developed in Aim 1 to drive advancements in pose-prediction and affinity prediction/ranking methods. Use feedback from DBP collaborators to improve the challenge technologies. Recruit new DBP partners on an ongoing basis.

Aim 3 Collaborate with and recruit DBP partners utilizing the end-to-end workflows from Aims 1 and 2 in drug-discovery application projects. Help researchers choose methods, based on their needs and the accuracy and efficiency of the methods. (**TRDP 3**.) Set up web-accessible servers for selected workflows. Collaborate with teams and Vanderbilt University and Boehringer-Ingelheim to use selected workflows to score candidate ligands designed by citizen-scientists.

Aim 4 Engage and enrich the computer-aided drug design research community. Continue Grand Challenges, initially at least, to focus community attention on a regularly scheduled, highly visible event and to recruit new DBP collaborators from the methods development community. Coordinate with other community-centered efforts, such as MolSSI and BioSimSpace, to develop accepted approaches and standards and foster best practices in CADD software development. Host software hackathons to help DBP investigators prepare their methods for the CELPP and CELPP+ servers and hold in-person and virtual workshops to share results and ideas.

RESEARCH STRATEGY

A. Significance

A.1. Resource Concept and Overview

The Drug Design Data Resource (D3R) is a community-oriented initiative that collects and uses scientific data to test and advance the state of the art in computer aided drug design (CADD) technologies by holding community-wide blinded prediction challenges. D3R focuses primarily on the core CADD challenges of ligand pose prediction and ligand affinity prediction or ranking, so key data types are high-quality ligand-protein cocrystal structures and measured ligand-protein affinities. From 2014 to the present, D3R has been funded as the sole NIH Drug Docking and Screening Data Resource under a NIH U01 cooperative agreement expiring at the end of September 2019. In this time, Professors Rommie E. Amaro and Michael K. Gilson, together with Professor Stephen Burley at the Research Collaboratory for Structural Biology Protein Data Bank (RCSB PDB)[1], have built D3R into a well-networked, community-centered, high-visibility international hub for community challenges in pose-prediction and ligand-ranking, and the challenges held by D3R have allowed some broad conclusions (below) to be drawn about the characteristics of successful CADD methods.

However, it has become clear that the standard approach of holding annual blinded prediction challenges with tens of crystal structures and several hundred affinities is too weak to advance the field further. Although such challenges are motivating and informative, they do not have the statistical power to distinguish clearly between most methods or to let a developer know whether his or her new approach is truly an improvement. At the same time, there is if anything a still greater need for technologies to rigorously and objectively assess CADD methods, as methods and applications continue to proliferate. **Accordingly, there is a clear recognition in the research community that a new approach to validation is required to realize the full potential of predictive computational methods for ligand discovery and design.** In particular, we need a way to dramatically increase the scale and throughput of, as well as the rigor and reproducibility of the methods employed in, blinded prediction challenges, so that detailed, informative questions about what works and what does not can be answered with confidence. This is the central goal of the present Biomedical Technology Research Resource (BTRR) P41 project described herein.

This project will build on the knowledge we have gained over the first 4 years as the D3R resource, coupled with the high degree of community engagement already established, and our technical depth in CADD and related data and computing technologies. Our technology development plan is organized into three Technology Research and Development Projects (TRDPs) that address CADD methods for pose prediction and ligand affinity prediction/ranking, and curate, archive, and disseminate methods and associated data. Integrated with the technology development plan is an extensive Community Engagement program, which includes a number of ongoing and newly proposed Driving Biomedical Research Projects with outstanding CADD method developers and scientists across the US and worldwide.

A.2. Background and Resource Vision

Scientific Background. The discovery of a small molecule that binds a disease-related protein with high affinity is a key step in many drug discovery projects. In the pharmaceutical industry, this step has been estimated to require over three years of work on average, at a net cost per launched drug rivaling that of clinical trials[2]. When a high-resolution structure of the targeted protein is available, a class of methods generally known as structure-based design methods, may be used to accelerate the discovery of high affinity ligands[3,4]. The computational challenges associated with structure-based ligand design generally comprise two main components. The first is to predict the bound conformation, or pose, of a candidate ligand, typically by fast, ligand-protein docking algorithms[5]. This information is of high value in its own right, as it can inform the designer of parts of the ligands that may be modified to form new interactions with the protein that seem likely to generate increased affinity. An accurate pose can also be used as a starting point for calculations aimed at estimating the candidate ligand's binding affinity for the targeted protein, or at least ranking its affinity relative to the affinities of other compounds one contemplates purchasing or synthesizing. When the structure of the target protein is not available, other computational methods of selecting high-affinity can be employed. For example, ligand-based methods[6] seek to identify or abstract key features or descriptors of ligands already known to bind the target and use these to suggest additional ligands that may bind with greater affinity. Recently, machine-learning methods that draw on wider datasets available in public resources such as BindingDB[7] and ChEMBL[8] have also been developed, and there are also computational methods that combine structure-based modeling with empirical learning[9].

Methods of pose-prediction and affinity prediction have been the subject of intensive research and development in both academic and commercial settings[10–12]. Nonetheless, the resulting computational methods have yet to fulfill their perceived promise, as neither is yet fully reliable[10–12]. In fact, as we have learned over the past 4 years running D3R, and as detailed below and in other parts of this proposal, it is surprisingly difficult even to compare the reliability of various methods in a consistent manner, and this limitation makes it correspondingly difficult to make and verify technical progress as new methods are devised.

Part of the challenge of rigorously comparing methods relates to the reproducibility (or lack thereof) of the complicated and highly variable end-to-end computational procedures for pose and affinity prediction[13]. Accordingly, we have seen a sharp rise in community interest in self-contained, automated workflows that clearly memorialize a particular procedure and provide defined approaches for their execution and deployment[14].

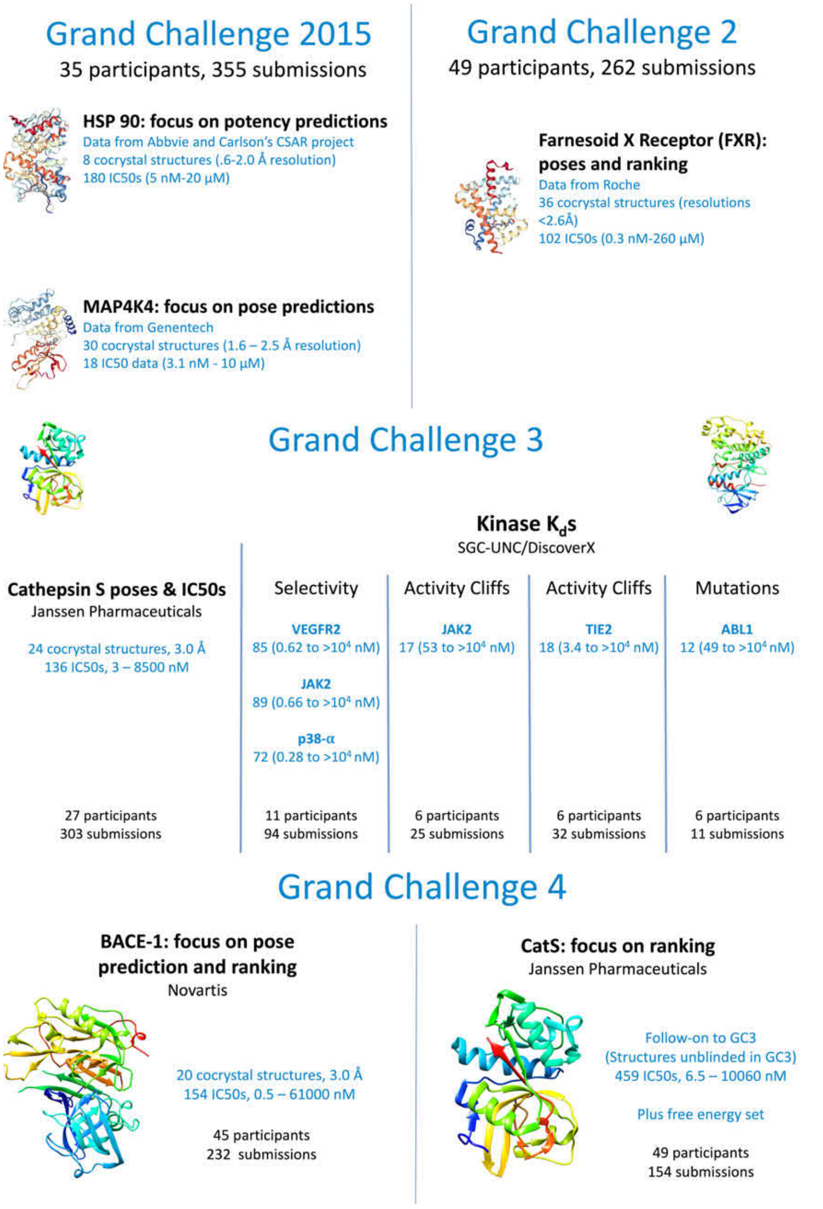


Figure 1. Summary of the D3R's annual Grand Challenges since late 2014.

In addition, although many performance comparisons have been published, the results can be difficult to interpret[15]. For example, new docking algorithms are frequently published along with a comparison against existing methods, but this comparison is often secondary to the description of the new algorithm, and hence not fully developed. Additionally, different methods are typically tested against different sets of protein-ligand complexes, so a consistent set of comparisons may not be available. Finally, even when a study carries out careful benchmarking of multiple methods against a common dataset, the dataset often comprises cocrystal structures and measured affinities that had already been published. Such retrospective studies are suboptimal, because they risk unintentional bias and because structures in the test set might have been used previously in training the docking algorithms[16]. Several initiatives have sought to address these limitations by hosting prospective, or blinded, prediction challenges. In such challenges, researchers evaluate methods against a common set of test cases for which the experimental structures are withheld until after the computational predictions have been made. Prior blinded challenges include the GSK challenge[15] and CSAR[17–19]. Similarly, in recent years, the Drug Design Resource (D3R) has run blinded prediction challenges called the Grand Challenges[10–12]; in fact, acquiring and curating the datasets and running the blinded challenges have constituted the primary goals of D3R under the existing NIH U01 award.

Thus, in the Grand Challenges, which we have run every year since our inception, D3R acquires high-quality X-ray crystal structures and additional binding affinity data (K_D's or IC50s) from, primarily, industrial pharmaceutical partners that have large internal databases of unpublished ligand-protein interaction data. We currently have active Data Transfer agreements with a number of companies, including Janssen, Roche, Glaxo Smith Kline, Genentech, Novartis, and Boehringer Ingelheim. Through these arrangements, we have brokered the acquisition of roughly a dozen targets and associated data and then used many of those data in the four Grand Challenges held to date[10–12]. These Grand Challenges have allowed method developers and CADD practitioners to test their methods and practices on typically 1-4 targets per year, each associated with tens of co-crystal structures for pose-prediction and a hundred to several hundred affinity data (**Figure 1**).

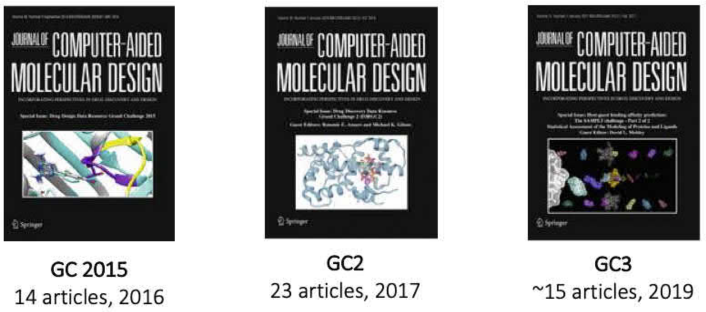


Figure 2. D3R's annual special issue of Grand Challenge reports, organized in JCAMD, are highly cited and provide the community with an effective mechanism to disseminate their Grand Challenge-related findings.

These efforts have been useful. They have highlighted the importance of rigorous method-testing, built a community of interest around rigorous evaluation methods, stimulated development of useful benchmarking strategies, provided some insights regarding best practices, and sometimes yielded unexpected results regarding the effectiveness of various technical approaches[10–12] (see below). The results are described in yearly Special Issues of the Journal of Computer-Aided Molecular Design **(Figure 2**), each of which contains a Grand Challenge overview paper from our team along with more detailed papers from most or all of the participants. These papers are well cited; thus, our 2015 and 2016 overview papers now have 62 and 29 citations, respectively, while the Grand Challenge 3 paper, which was published about a month ago (10 Jan 2019) already has 122 downloads.

As noted above, the Grand Challenges have generated some broad conclusions about pose- and affinity-prediction or -ranking methods; notably:

- Successful prediction of ligand-protein poses depends on the entire workflow, including factors extrinsic to the core docking algorithm, such as how the ligand and protein structures are prepared, the conformation of the protein selected, and the treatment of crystallographic waters.
- The success of docking and scoring predictions is not clearly correlated with the software used.
- Using existing structural information, such as cocrystal structures of small molecules with the target protein, can increase docking success rates. For example, known poses of similar ligands can guide positioning of the new ligand, and better docking results may be obtained by docking a new ligand into a binding site solved with another ligand with the same chemotype.
- There is a mixed picture as to whether human intervention leads to improved results.
- The accuracy of the poses used in structure-based affinity rankings does not clearly correlate with ranking accuracy.
- At least in these studies, explicit solvent free energy methods have not yet outperformed faster scoring methods.

However, there are still many important questions the Grand Challenges have not been able to answer. For example, it has not been possible to draw truly rigorous conclusions about whether particular methods provide statistically significantly higher accuracy, or to determine whether some methods work best for certain types of target. Thus, the blinded prediction challenges to date have not been large and systematic enough to afford statistically meaningful distinctions among individual methods or to support an efficient cycle of development and evaluation that can persistently accelerate progress in the field. In addition, the prediction challenges to date have not overcome the problems associated with fully documenting CADD methods and replicating results.

Resource Vision. In order to overcome the limitations of standard prediction challenges, D3R proposes to develop an integrated set of technologies that will enable us to challenge CADD methods with a dramatically greater flow of test cases, rising from today's tens of pose-predictions per year to thousands per year, and from today's hundreds of affinities per year to tens of thousands per year. As part of this effort, we will work with DBP investigators to help capture their CADD methods in end-to-end, containerized workflows. These will not only enable the automation required to handle the high volume of calculations but will also take reproducibility and dissemination to a new level. Thus, a key output of this work will be a unique new archive containing fully operational workflows along with the experimental data used to evaluate them and the results of their predictions of these data. These technologies and the data they generate will take the testing of CADD methods to a new level, providing DBP developers with a powerful new tool to improve their methods, and providing DBP drug designers with the ability to choose methods best suited to their projects based on solid test results. The present

BTRR project will build on D3R's extensive experience with holding blinded prediction challenges and our visibility in the CADD community, as well as our team members' proven track record in data curation and management, database development, structural biology, and workflow and containerization technologies.

Preliminary results. As a proof-of-principle and illustration of the new approach, we have piloted a high-throughput, blinded pose-prediction challenge called Continuous Evaluation of Ligand Protein Predictions, or CELPP[20]. The CELPP challenge takes advantage of the fact that the PDB releases a new set of structures every week, and a number of these are co-crystal structures of a protein with a drug-like molecule and hence suitable for a pose-prediction challenge. D3R's structural biology Lead Prof. Stephen Burley (Director of the RCSB PDB) worked with the other PDB sites to develop a procedure in which a file listing all forthcoming entries, including their PDB ID, protein sequence(s), InChIs of any ligands, and the pH of the crystallographic mother liquor, is released several days before the structures themselves. Each week, in-house CELPP scripts download this file and identifies those which contain protein-small molecule cocrystal structures suitable for automated docking calculations (https://github.com/drugdata/D3R). Most weeks, there are 20-50 proteins co-crystallized with suitable ligands, recently averaging ~27 heavy atoms and 5 rotatable bonds. Pose predictions are carried out using our own D3R-developed in-house workflows, as well as by a handful of early adopters (now **DBP Investigators 1, 8, and 16**), who have each automated their own docking codes to handle this continuous challenge format. This rolling challenge was inspired by the Continuous Automated Model Evaluation (CAMEO) protein structure prediction challenge[21,22], which also takes advantage of the PDB's weekly structure release.

In slightly over a year (62 weeks), CELPP provided 3,184 blinded ligand-target co-complexes for docking. **This number surpasses by an order of magnitude all of the pose-prediction cases presented in community-organized blinded challenges over the past decade combined,** including D3R[10–12], CSAR[17–19] and GPCR Dock[16], which in total used about 320 structures. Thus, by taking advantage of existing data in a new way, we have created the possibility of generating statistically significant results for pose prediction methods. For this preliminary study, we analyzed results for the first ~1600 structures, based on RMSD (Å) from the crystallographic pose. It was perhaps surprising that, as detailed elsewhere[20], the fully automated CELPP procedures did about as well overall as the Grand Challenge participants, given that the latter were not necessarily automated. Interestingly, most of the docking methods tested exhibited rather similar levels of accuracy (**Figure *3***), based on median RMSD, with rDock[23] and one External Participant performing noticeably worse. Among our in-house workflows, OE Fred[24–26] and Vina[27] yield somewhat lower RMSDs, with GLIDE[28–30] and rDock[23] trailing slightly by this metric. Note that these workflows are not tuned for optimum results so these results may not reflect the best performance available from their respective algorithms.

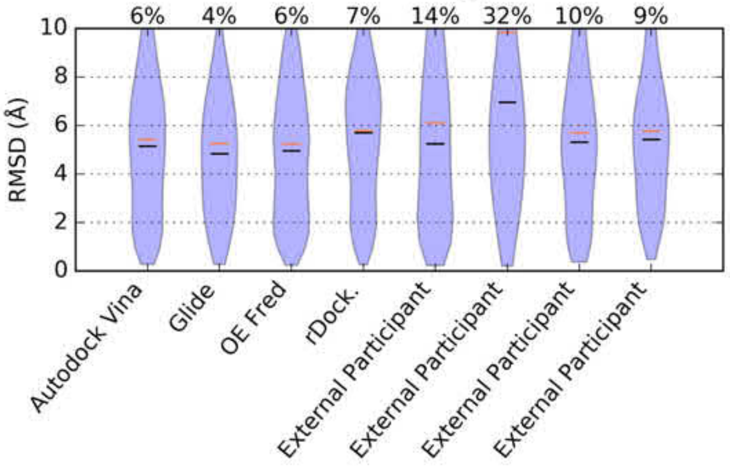


Figure 3. Violin plots showing the distributions of pose errors for various docking methods implemented by D3R and by four external participants, as labeled, for ~1600 test cases. Black line: median. Orange line: mean. Above each plot is the percent of predictions with RMSD>10Å.

Overview of technologies to be developed and DBPs to be supported. We aim to further develop the initial CELPP[20] method described above, and also to create an entirely new analog of CELPP, tentatively called CELPP+, which will provide high-throughput automation of protein-ligand affinity predictions or rankings.

First, we will convert the initial CELPP technology into a more robust service for the community of docking software method developers and practical users (Technology Research & Development Project 1; **TRDP 1**), by developing a Kubernetes-based server that will host fully-automated, containerized workflows for pose prediction, created by D3R and **DBP** collaborators (**Figure 4**). This will be connected to the PDB pre-release system, and also to a new D3R database (D3RDB), which will archive and disseminate the resulting predictions, workflows, and associated provenance; i.e., other required parameters, such as library dependencies and hardware details that are required to faithfully reproduce the experiment (**TRDP 3**). This archive will also be used by drug-design **DBPs 7 and 15** looking to choose the most appropriate pose-prediction methods for their projects.

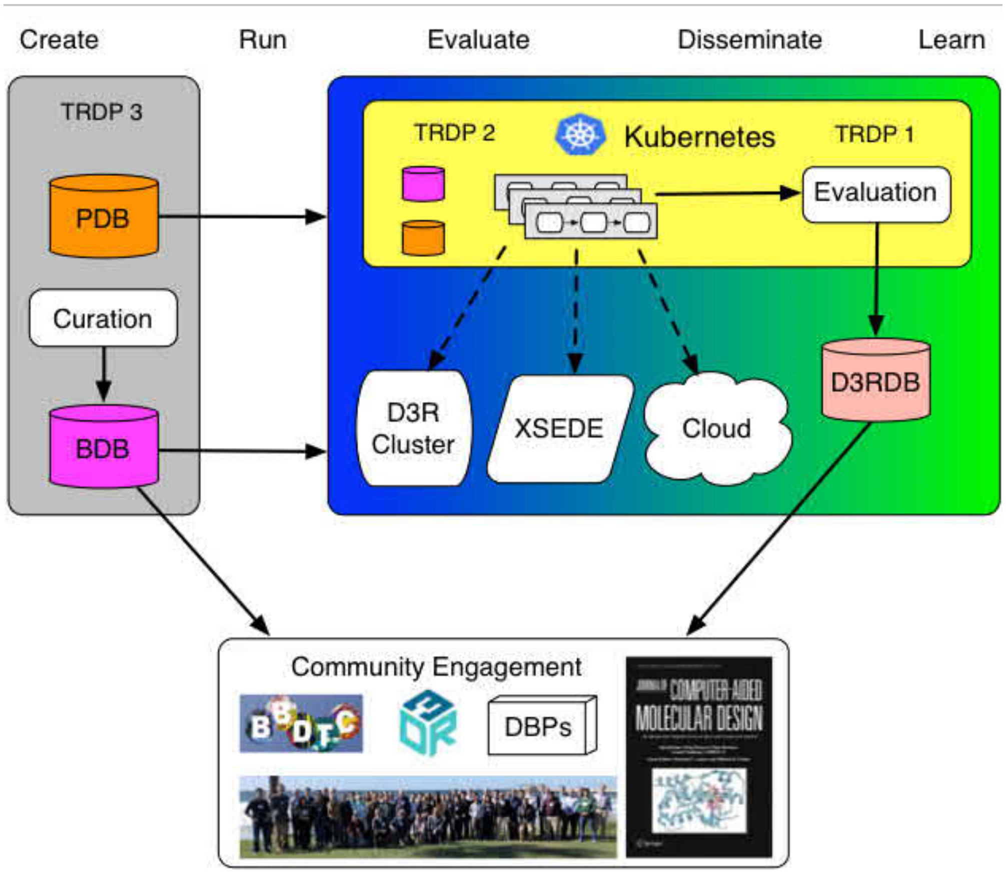


Figure 4. High-level diagram of relationships among project components.

To create CELPP+, we will set up a server (using the Google-developed, now open source, broadly adopted and supported Kubernetes[31] framework) hosting affinity-prediction workflows, which will be time-stamped according to when they were provided to D3R, and which will be blocked from accessing data resources on the internet. We will then adopt and direct the data flow from the Gilson lab's existing high-throughput curation of protein-ligand affinity data, which adds tens of thousands of data per year to the BindingDB database[7], to the workflows in the cluster, thus challenging them to predict results which became publicly accessible only after the date the workflows were commissioned. This novel strategy will generate the first effectively blinded, continuous challenge for protein-ligand affinity prediction or ranking, because once the containerized workflows are brought to our server, they can no longer be tweaked by the developers. Again, the workflows, binding data, and prediction results will be archived in D3RDB, which will be integrated for this purpose with BindingDB[7], and made available to **DBP** investigators and the rest of the research community.

B. Innovation

The technologies to be developed here are new and have high potential to create a unique and valuable new resource. Indeed, until now, no one has proposed a strong and viable solution to resolve the major impediments to CADD method development discussed above. As proven by our preliminary implementation of CELPP, there are important opportunities to take advantage of existing data streams to test and improve CADD methods. If successful, this pioneering technology development approach stands to revolutionize methods development in the prediction of ligand-protein poses and affinities. In addition, recent advances in workflow and containerization technologies will support replication, dissemination, and further development of the methods tested in this system. Thus, our efforts will uniquely identify and foster best practices for use by the many users of **DBP** investigators' methods, which we estimate as at least >30,000 researchers worldwide, based on just three of the docking codes (i.e., **DBPs 2, 5, and 12**).

C. Approach

C.1 Technology Research & Development Projects

To effectively carry out this work, we organize our efforts into three TR&D Projects (hereafter, TRDPs), as now summarized.

TRDP 1: Pose Prediction. The efforts of this project center on advancing pose prediction methods by converting CELPP to a workflow-based challenge (Aim 1), developing and deploying an advanced website to share CELPP results and workflows (Aim 2), and analyzing CELPP results to extract scientific insights and value (Aim 3)**.** The main data flows and operations are diagrammed in **Figure 5**.

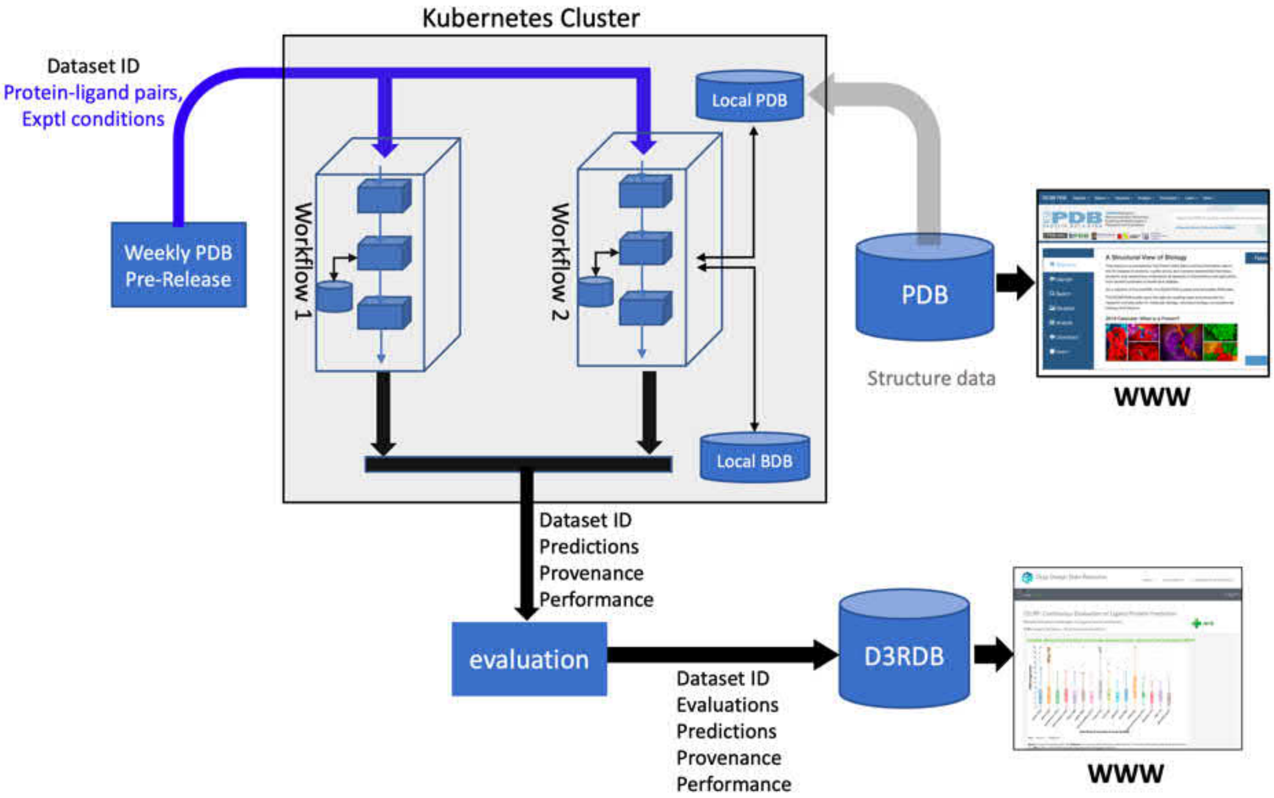


Figure 5. In TRDP 1, CELPP will take pre-release data from the PDB, feed it into containerized workflows in our Kubernetes environment, evaluate, archive, and disseminate results, evaluations, provenance, compute performance, and workflows. Two workflows are shown, each comprising three functional modules and having a persistent data store. Workflow 2 is shown as using the local PDB and BindingDB instances.

TRDP 2 – Ligand Ranking and Affinity Predictions. This project will deliver a novel, effectively blinded, high-throughput challenge for protein-ligand affinity prediction or ranking, preliminarily named CELPP+, after the CELPP pose-prediction challenge described in **TRDP 1**. We will develop the CELPP+ workflow framework and standards for affinity prediction workflows (Aim 1), populate the CELPP+ framework with affinity-prediction workflows (Aim 2), collect, archive, and broadly share results and workflows (Aim 3), and, finally, extract insights and value from the assessments (Aim 4). The main data flows and operations are diagrammed in **Figure 6**.

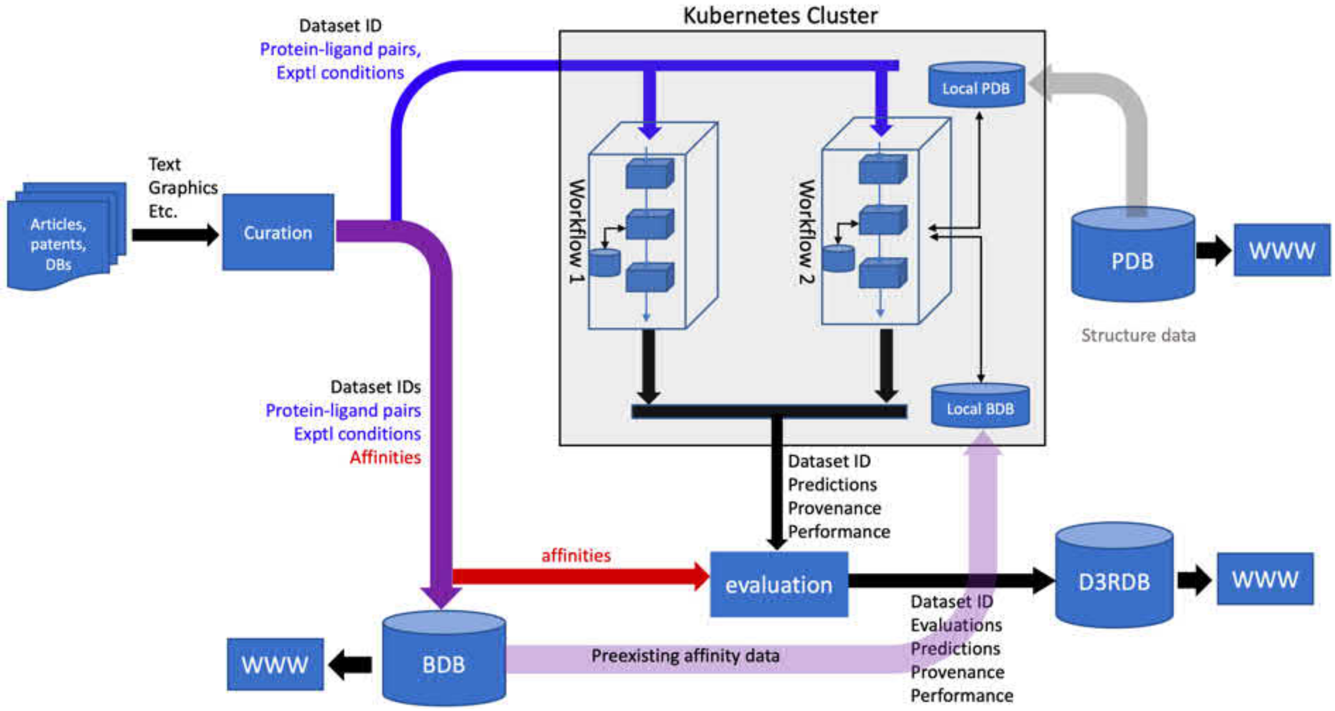


Figure 6. In TRDP 2, BindingDB curators will collect affinity data and route it to BindingDB (BDB) for storage and dissemination. In parallel, the identities of the proteins and ligands will be sent to the Kubernetes cluster where workflows will predict the affinities. The evaluator compares with predictions with the experimental data and completes a data package of results which is routed to D3RDB, which is integrated with BindingDB (integration not shown). The workflows are modular and have access to local instances of BindingDB and the PDB for their calculations.

TRDP 3 – Data and Analytics. This project will provide the information infrastructure to capture, archive, and disseminate all of the methods and data developed in the course of this project. We will collect and curate protein-ligand interaction data (Aim 1), archive workflows and challenge results with associated protein-ligand interaction data (Aim 2), and serve data to project websites, CELPP+ workflows, and the research community (Aim 3). The main data flows and operations are diagrammed in **Figure 7**.

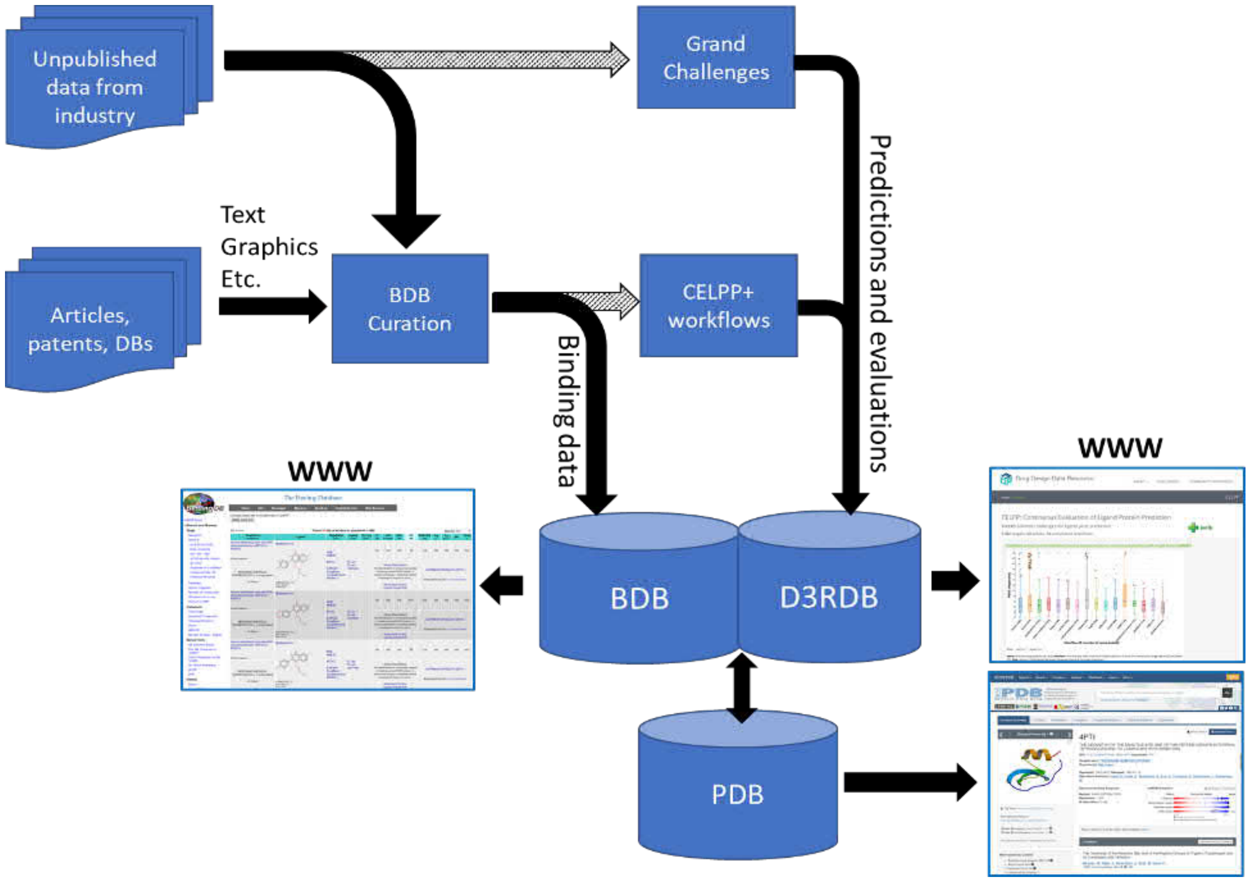


Figure 7. Data flows for affinity prediction challenges. Hatched arrows indicate challenge data "blinded" by removal of the experimental affinities and simplified by removal of needless ancillary information, such as literature citations. BindingDB (BDB) and the D3RDB of challenge methods and results are shown as fused, to indicate that D3RDB will be constructed as an integrated extension of BindingDB.

C.2 Overall Themes of Driving Biomedical Project (DBP) Portfolio

Each of our TRDPs supports and is driven by a set of DBPs. To provide a high-level perspective and succinctly describe how our activities are driven by various stakeholder needs, we have organized our main initial DBP portfolio into three broad Aims, as follows.

Aim 1: Enabling Evaluation and Advancement of Pose Prediction. Here, labs that develop pose-prediction, or docking, technologies, will encapsulate their methods in automated workflows and place them in the CELPP+ data stream on the D3R server. This will provide them with a continuous flow of evaluative data, which they will use to test and improve their methods. Thus, our work will serve the large community of pose prediction (docking) method developers, and, by extension, the many users of these codes. Our initial list of DBPs in this category includes the following well-established methods / code bases: AutoDock[32,33] and AutoDock VINA[27] (> 29,000 registered users worldwide, **DBP 5**); DOCK[34] (est. > 10,000 registered users worldwide, **DBP 12**); HADDOCK[35] (> 12,000 registered users worldwide, **DBP 2**); MolSoft[36] (**DBP 13**); Rhodium (**DBP 1**); SMINA[37] (**DBP 8**) and MathPharm (**DBP 15**). At the same time, feedback from these developers will help us streamline and improve CELPP.

Aim 2: Enabling Evaluation and Advancement of Ligand Affinity Prediction or Ranking. Here, similarly, labs that develop affinity prediction or ranking methods will use CELPP+ to test and improve their methods. Many of the same investigators who will use CELPP will also use CELPP+. In addition, CELPP+ will serve scientists who focus more on affinity prediction than pose prediction, such as the developers of AIGDock[38] (David Minh, **DBP 11**), free energy method developers[39,40] (Zoe Cournia, **DBP 4** and Julien Michel, **DBP 10**), and potentially ligand-based and pure machine-learning methods. Again, we will use feedback from the DBP investigators to drive improvements in CELPP+.

Aim 3: Applications of Evaluated CADD Methods & Innovative Uses of D3R Technologies. The third group of DBP investigators are those who will select and use D3R-evaluated workflows, created by DBP investigators or D3R, for their drug discovery projects. Here, we have limited our DBPs to a small handful we view as exceptionally impactful or innovative uses of our developing technologies, though we note that we expect our workflows and ultimately, the web services, to be of enormous use in biomedical research programs. The DBPs under this aim include projects targeting Alzheimer's disease (**DBP 14**), and cancer and AIDS (**DBP 6**). It is also worth highlighting a particularly innovative project (**DBP 9**) to create citizen-scientist ligand-design platform (**Meiler Letter**) analogous to the FoldIt protein design game[41], in collaboration with researchers at Boehringer Ingelheim (**Bergner Letter**), as diagrammed in **Figure 8.** Here, D3R's role will be to provide one or a few evaluated workflows, accessible by webservices, as a second-level scoring function for promising ligands. Note that this platform will ultimately be available for use by drug designers at any university or drug company.

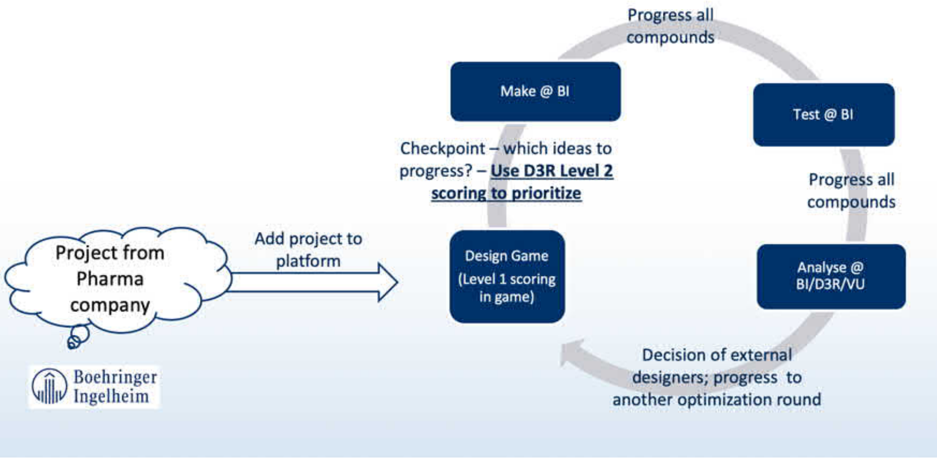

Progress all compounds
Make @ BI
Test @ BI
Checkpoint – which ideas to progress? – Use D3R Level 2 scoring to prioritize
Progress all compounds
Project from Pharma company
Add project to platform
Design Game (Level 1 scoring in game)
Analyse @ BI/D3R/VU
Decision of external designers; progress to another optimization round
Boehringer Ingelheim

Figure 8: Crowd-sourced ligand design in DBP with Meiler (Vanderbilt) and Bergner (Boehringer Ingelheim, BI). Initial sketch of the process is as follows. Pharma company, here BI, adds project, with protein target structure, to the game. Citizen scientists build and design candidate ligands using Level 1 in-game scores. Top scoring candidates are rescored with Level 2 scoring using top-performing D3R workflows. The top compounds are then synthesized and tested at the pharma company. Analysis occurs across the three sites and then, if desired, additional rounds of optimization will ensue.

One added **TRDP3** collaborator (see **Letter of Leon Bergen** in **TRDP 3**) will use our corpus of curated data to train machine learning and NLP techniques from linguistics and language processing to improve the efficiency of the curation process that drives CELPP+.

C.3 Projected Arc of Technology Research and Development

Based on our knowledge of the field and our experience over the past 4 years establishing D3R as an NIH U01 project, we expect that this project will be productive for 10-15 years as a NIH P41. Here we lay out our long-term vision for the Drug Design Data Resource.

During the first 5 years as a P41 BTRR, D3R will establish robust initial workflow-based technology frameworks for both CELPP and CELPP+. We also anticipate continuing our roughly annual Grand Challenges for at least the first few years, as a means of interaction with the community and recruiting new DBPs. We will help lead the national and international conversations around data standards and methods standards for CADD, along with other entities, including the NSF-sponsored MolSSI Institute (see **Letter of Daniel Crawford**) the UK's flagship simulation software project, BioSimSpace (see **Letter of Julien Michel**), and we will follow these standards in our operations. We will work with community members (e.g., DBP Investigators) to develop containerized, modular workflows for docking and affinity prediction and will develop the best performing open source workflows into accessible web services. We will recruit new DBP collaborators across all three Aims, adapting as needed in response to changes in the field.

In years 6-10, we anticipate operating in full production mode, accommodating many new workflows by migrating much of the compute load from our cluster to cloud resources. More work may need to be done breaking workflows into interoperable modules, but this will enable complex experiments that integrate methods from different groups into hybrid end-to-end workflows and lead to improved methods. These will support a growing community of data-science-embracing DBP users with drug-design applications. We anticipate that, over time, advanced computational methods developed in the DBPs using D3R technologies will be broadly and that their use, maintenance, and further development will be sustained by individual research grants and/or through commercialization. The CELPP and CELPP+ frameworks will also start becoming robust enough that some other groups may begin installing them on their own systems for in-house research. Note, however, that these will still depend on our data curation in **TRDP 3**. Thus, we will develop opportunities to control costs, such as reducing data curation costs and fostering sustainability by working with journals to require direct deposition of protein-ligand binding data into the public databases, and by collaborating with machine learning and natural language processing experts to ramp up reliable methods of partially or fully automating the extraction of benchmark data from articles and other documents—as indeed already begun in the current proposal. Finally, given community interest, it may be useful to take analogous approaches (including acquiring relevant data from industry) to address other outstanding challenges in CADD, such as predicting drug binding kinetics, membrane permeability, or toxicity.

Toward the end of years 6-10, we will assess the future role of D3R as then constituted. Much may have changed in the fields of CADD and software engineering, but we anticipate new opportunities to enhance CELPP and CELPP+ by integrating new software technologies, and to recruit DBP collaborators with innovative prediction methods. For example, assuming continued improvement in computer speed, we anticipate growing use of simulation-based free energy methods, possibly married to advanced machine-learning technologies. More exotic possibilities may also merit consideration, such as integrating our technologies with robotic chemical synthesis and testing, leading to integrated, iterative, self-learning structures that can evolve not only computer models but also targeted compounds with minimal human intervention.

C.4 Approach to Community Engagement

Serving the CADD community. D3R has been and will continue to be an intensely community-centered effort. The scientific community that D3R consists of several different stakeholder groups. First and foremost, we serve the CADD method developers; these are the teams that develop new methods for pose prediction and ligand ranking or binding affinity. The methods developed by these teams see wide use in the biomedical research enterprise, with a number of our DBP investigators having over 10,000 registered users worldwide; e.g., AutoDock Vina[27], HADDOCK[35], and DOCK[34]. D3R also serves the industrial CADD community, expert users within pharmaceutical and biotech companies (see **Letters of Support from Tara Mirzadegan (Janssen), Richard Lewis (Novartis)**, **Georgia McGaughey (Vertex)**, **Woody Sherman (Silicon Therapeutics);** and in Admin**: Pat Walters (Relay Therapeutics), Hanneke Jansen (Novartis), Martin Stahl (Roche)**; and in DBPs: **Gruson & Steuer on behalf of Andreas Bergner (Boehringer Ingelheim)**. These individuals also have enabled the D3R Grand Challenges[10–12] by releasing hitherto unpublished datasets to D3R. The shared engagement of the CADD community brings substantial impact to the research enterprise and overall public health.

Figure 9. Members of the community we serve attending the 2018 D3R workshop in La Jolla.

As a central hub for CADD developers worldwide, D3R maintains an exceptionally high level of visibility, and we have already established a solid track record of productive community engagement. Our approach to community engagement has been multifaceted. In addition to organizing and running the Grand Challenges[10–12], we also developed the CELPP rolling challenge discussed above, hosted workshops, published in peer-reviewed journals, and participated in high profile conferences. We also use D3R to engage the community in intensive discussions around important dimensions of the field. For example, at our last workshop in 2018, after a presentation from our external evaluator Pat Walters (from Relay Therapeutics, see **Letter of Support in Admin**) about the longitudinal findings of our Grand Challenges to date, the community came to consensus about the need to develop a working group to develop a set of evaluation metrics to be applied for different challenge categories. Thus, there is substantial opportunity to leverage community engagement activities to enrich our resource activities. This includes opportunities to develop CADD standards, spanning all aspects of the end-to-end process, including data i/o standards, process/ methods standards, evaluation standards, and publication standards. We plan to continue all of the aforementioned activities, and also to engage in multiple dimensions of training activities, including developing online training materials for training of a more general audience (which we expect will increase our visibility further), as well as hosting intensive hackathons for our DBP investigators and other interested parties, to increase use, functionality, and adoption of our developing technologies. The following subsections touch on key aspects of our plan for community engagement.

Blinded challenges. The annual Grand Challenges (GCs) have been a cornerstone of our activities[10–12]. Participation has grown from 37 participants in GC2015 to 55 participants in our most recent GC4, and our most recent challenge (GC4) drew worldwide participation (**Figure 10**). We plan to continue these and use them as a central point of our in-person workshops. These will be supplemented by the new, continuous CELPP and CELPP+ challenges proposed here.

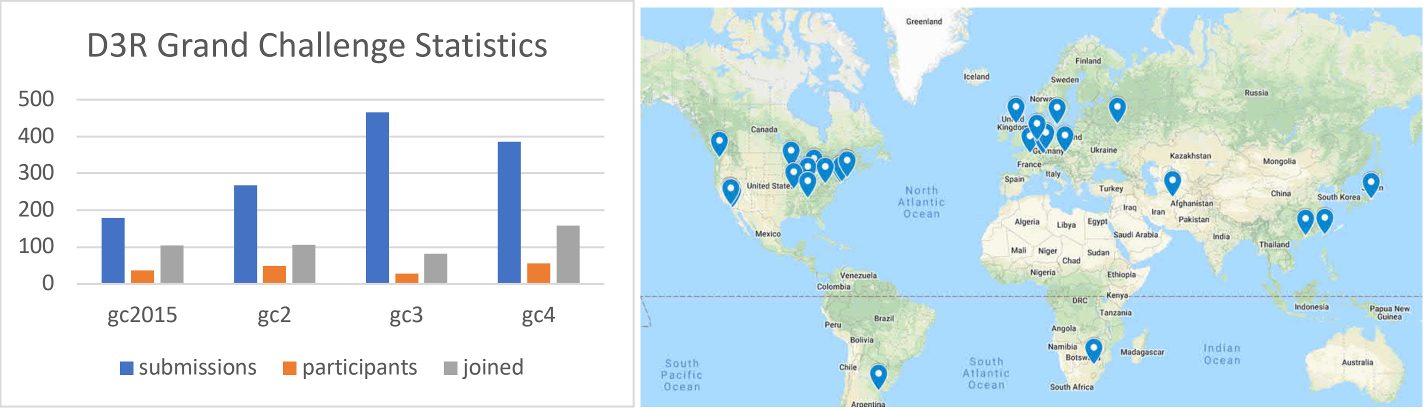


Figure 10. Left: Grand Challenge submissions (blue), participants (red), and number of people who 'joined' the challenge (gray). Right Global participation in D3R's Grand Challenge 4.

Workshops and hackathons. Our in-person workshops in La Jolla on the UC San Diego campus have been well-attended, with typically about 70-90 in-person and an equal number of remote participants, and we will continue to hold these along with virtual workshops, for an annual meeting schedule. We also extended the reach of our community engagement by partnering with the SAMPL initiative (see **Letter of Support from David Mobley**), which focuses on model systems for method development[42,43]. By intersecting these two groups, D3R will continue to enable a highly productive interchange of ideas. We also plan to host intensive hackathons to help DBP Investigators and other interested researchers develop their methods into end-to-end workflows for CELPP and/or CELPP+ and to containerize them so that they can be mounted on our Kubernetes[31] servers and participate in these new challenges.

Training. During our interactions with D3R industry partners, we often receive feedback about training gaps in the rising workforce, which is largely trained by academic researchers. We will take advantage of our central position in the CADD community to identify key training areas and will identify trusted community members to serve as, for example, instructors at workshops. Many CADD scientists in the industrial section are eager to share their knowledge with the next generation of trainees and they will represent a valuable but untapped source of expertise and career perspectives. These training events will be used as the basis to create on-line training modules which will be made freely available at the D3R website and beyond. We also believe it is important to introduce trainees to data science aspects of CADD, and will utilize our innovative, collaborative training platform, the Biomedical Big Data Training Collaborative (BBDTC), as part of these efforts. Thus, in 2015, D3R investigators Altintas and Amaro were awarded an NIH BD2K R25 Open Educational Resource grant to establish the BBDTC, an online learning platform that includes big data science curricula, an open online course (OOC) framework, and a software toolbox for assembling hands-on tutorials and executing containerized workflows.

Publishing D3R Results in Peer-Reviewed Journals. Perhaps the most effective way to communicate to the scientific community is through peer-reviewed journals. Accordingly, D3R has ensured that our findings and those of our participants are shared via the scientific literature. Since 2015, Amaro and Gilson have co-edited special issues of the Journal of Computer Aided Molecular Design (JCAMD) reporting on the most recent Grand

Challenge. In addition to D3R's own overview paper, which summarizes the performance of all of our participants and distills broad conclusions, these issues also includes papers by most or all of the GC participants[10–12]. We plan to continue these well-received publications (see **Letter of Terry Stouch**).

Disseminating information through the D3R website, email and Twitter. We maintain a comprehensive D3R website (https://drugdesigndata.org) developed on the SciCrunch platform. Our website, which conforms to the P41 specifications provided in the PAR, has proven to be a highly effective communications vehicle with steadily increasing user visits since its launch in 2015. It currently provides current challenge information, archives of prior challenges, links to special issues and workshop information, and prototype CELPP webpages, which will be greatly extended as CELPP and CELPP+ become operational. Challenge participants are registered into a challenge-specific email list, which allows focused communications on each evolving challenge. In addition, we maintain a Twitter presence to expand D3Rs' exposure to interested researchers. These activities, too, will continue as we move to the P41 format.

Disseminating Data, Scripts and Software Products. D3R software is maintained as open-source to the extent possible and is shared on the widely used GitHub repository. We will continue this practice in the present project. In addition, as detailed in the Resource Sharing section and the Data and Analytics TRDP, curated data and detailed challenge results will be archived and shared via the D3RDB/BindingDB and PDB databases.

D. Research Strategy Annotated Timeline. The following outline sketches key steps in the project timeline. Some ongoing or repetitive items are not listed, such as Grand Challenges, data curation, basic website development, and consultations with MolSSI, BioSimSpace and others.

- Year 1
  - Cluster purchase and setup
  - Prototyping of containerized versions of D3R's existing CELPP and CELPP+ workflows
  - Initial port and testing of CELPP scripts to Kubernetes framework
  - Design of CELPP+ system, including data transfer standards
  - Design and initial implementation of D3RDB/BindingDB data dictionaries and database extensions
- Year 2
  - Operation, data generation, and ongoing refinement of initial CELPP system
  - Development of automated flows of data from and into the D3RDB/BindingDB database archive
  - Partial implementation and initial testing of CELPP+ system in Kubernetes framework
  - Deployment, testing, and refinement of first DBP-generated workflows
  - Development of middleware between D3RDB/BindingDB and webpages
- Year 3
  - Full scale operation, data generation, and ongoing refinement of CELPP and CELPP+ with D3R and DBP-generated workflows
  - Installation of new DBP CELPP and CELPP+ workflows
  - Development of standards and procedures for modularization of workflows
  - Initial use of CELPP and CELPP+ to study methodological determinants of success, run parameter scans, study ranges of applicability
  - Begin development of select workflow-based, web-accessible servers
  - Design and initial development of D3RDB/BindingDB query, browsing and download tools.
- Year 4
  - Operation, refinement and hardening of CELPP and CELPP+
  - Modularization of sample D3R workflows and refinement of modularization approach
  - Installation and evaluation of new DBP CELPP and CELPP+ workflows
  - Further studies of methodological determinants of success, parameter scans, ranges of applicability
  - Completion and deployment of selected web-accessible pose-prediction and affinity-ranking servers
  - Further development of D3RDB/BindingDB query, browsing and download capabilities
- Year 5
  - High-throughput CELPP and CELPP+ data generation and archiving
  - Development of advanced website capabilities using new D3RDB/BindingDB capabilities
  - Evaluation and analysis of all operational DBP workflows
  - Design and testing of new workflows based on insights from prior analytic studies.
  - Deployment of additional web-accessible servers
  - Demonstration and possible use of CELPP and CELPP+ on cloud resources

REFERENCES

1. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
2. Paul, S. M. *et al.* How to improve RD productivity: The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
3. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. 334–395 (2014).
4. Kuhn, B. *et al.* A Real-World Perspective on Molecular Design. *J. Med. Chem.* **59**, 4087–4102 (2016).
5. Sinko, W., Lindert, S. & McCammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **81**, 41–49 (2013).
6. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
7. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
8. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
9. Mura, C., Draizen, E. J. & Bourne, P. E. Structural biology meets data science: does anything change? *Curr. Opin. Struct. Biol.* **52**, 95–102 (2018).
10. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **30**, 651–668 (2016).
11. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* **0**, 0 (2019).
12. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* 1–20 (2017). doi:10.1007/s10822-017-0088-4
13. Jansen, J. M., Cornell, W., Tseng, Y. J. & Amaro, R. E. Teach–Discover–Treat (TDT): Collaborative computational drug discovery for neglected diseases. *J. Mol. Graph. Model.* **38**, 360–362 (2012).
14. Pronk, S. *et al.* Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform. *J. Chem. Theory Comput.* **11**, 2600–2608 (2015).
15. Warren, G. L. *et al.* A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **49**, 5912–5931 (2006).
16. Weiss, D. R., Bortolato, A., Tehan, B. & Mason, J. S. GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J. Chem. Inf. Model.* **56**, 642–651 (2016).
17. Damm-Ganamet, K. L., Smith, R. D., Dunbar, J. B., Stuckey, J. A. & Carlson, H. A. CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* **53**, 1853–1870 (2013).
18. Carlson, H. A. Lessons Learned over Four Benchmark Exercises from the Community Structure–Activity Resource. *J. Chem. Inf. Model.* **56**, 951–954 (2016).
19. Carlson, H. A. *et al.* CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **56**, 1063–1077 (2016).
20. Wagner, J. *et al.* Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *bioRxiv* (2018). doi:10.1101/469940
21. Haas, J. *et al.* Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins Struct. Funct. Bioinforma.* **86**, 387–398 (2018).
22. Haas, J. *et al.* The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database* **2013**, (2013).
23. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Comput. Biol.* **10**, e1003571 (2014).
24. Kelley, B. P., Brown, S. P., Warren, G. L. & Muchmore, S. W. POSIT: Flexible Shape-Guided Docking For Pose Prediction. *J. Chem. Inf. Model.* **55**, 1771–1780 (2015).
25. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **51**, 578–596 (2011).
26. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided Mol. Des.* **26**, 897–906 (2012).
27. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
28. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **49**, 6177–6196 (2006).

29. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
30. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
31. Burns, B., Grant, B., Oppenheimer, D., Brewer, E. & Wilkes, J. Borg, Omega, and Kubernetes. 24
32. Goodsell, D. S., Morris, G. M. & Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit. JMR* **9**, 1–5 (1996).
33. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
34. Allen, W. J. *et al.* DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* **36**, 1132–1156 (2015).
35. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
36. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506 (1994).
37. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
38. Xie, B. & Minh, D. D. L. Alchemical Grid Dock (AlGDock) calculations in the D3R Grand Challenge 3 : Binding free energies between flexible ligands and rigid receptors. *J. Comput. Aided Mol. Des.* (2018). doi:10.1007/s10822-018-0143-9
39. Athanasiou, C., Vasilakaki, S., Dellis, D. & Cournia, Z. Using physics-based pose predictions and free energy perturbation calculations to predict binding poses and relative binding affinities for FXR ligands in the D3R Grand Challenge 2. *J. Comput. Aided Mol. Des.* **32**, 21–44 (2018).
40. Mey, A. S. J. S., Jiménez, J. J. & Michel, J. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *J. Comput. Aided Mol. Des.* **32**, 199–210 (2018).
41. Kleffner, R. *et al.* Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics* **33**, 2765–2767 (2017).
42. Bannan, C. C. *et al.* Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **30**, 927–944 (2016).
43. Yin, J. *et al.* Overview of the SAMPL5 host–guest challenge: Are we doing better? *J. Comput. Aided Mol. Des.* **31**, 1–19 (2017).

TR&D Project 1: Pose Prediction

SPECIFIC AIMS

An accurate method to predict the bound conformation, or pose, of a small molecule in a protein's binding pocket would allow drug designers to quickly identify chemical sites on a ligand that are disposed in space in a manner that could direct a new substituent into a targeted subsite of the binding pocket, and thus potentially to design a compound with greater binding affinity. In addition, accurate pose prediction, or docking, is a necessary first step in structure-based methods to predict ligand-protein affinities. It can also provide valuable insight into biomolecular mechanisms. For example, knowing precisely how an enzyme binds its substrate is essential to understanding its catalytic mechanism. As a consequence, researchers worldwide are working to solve the so-called docking problem.

However, progress toward this goal is inhibited by the fact that it is difficult to determine whether a change in a docking method consistently yields more accurate predictions than prior methods, or to compare two distinct methods on a solid footing and thus to select one for a given application. One reason is that different methods are typically tested against different sets of protein-ligand complexes, so there is no consistent basis for comparison. Moreover, reported tests of docking methods typically are based on publicly available protein-ligand cocrystal structures, and many of the same publicly available test cases are used over and over. Such studies risk unintentional bias, due to knowledge of the reference data and because structures in the test set might have been used previously in training the docking algorithms.

The present TRDP aims to overcome these problems and establish a foundation for systematic improvement of docking methods. We will build on strong preliminary studies to establish a powerful new evaluation technology that will be made freely available for use by collaborating DBPs. The method, called Continuous Evaluation of Ligand Pose Predictions (CELPP), uses the PDB's weekly release of new protein-ligand cocrystal structures as the basis for a rolling, blinded pose-prediction challenge. The docking methods will be encapsulated in automated, containerized workflows hosted on the D3R server, and the evaluation results, coupled with the test data, will be archived along with the workflows themselves, in the database to be developed in TRDP 3. Key outcomes will include detailed evaluations of widely used docking methods; an ongoing evaluation method based on new data each week to drive improvements in docking technology; insights into the determinants of accuracy in docking methods; and an archive of evaluated, downloadable workflows for use in drug discovery projects.

This will be accomplished via the following specific aims:

Aim 1: Convert CELPP to a workflow-based challenge. In the existing, preliminary implementation of CELPP, the docking calculations are done by off-site research groups. We will work with these and other DBP investigators to package their codes into containerized, modular workflows which will be run on a Kubernetes framework on the D3R server. Benefits will include collection of statistics on stable methods, automated archiving of results via TRDP 3, and support for fine-grained reproducibility and facile dissemination and reuse of methods.

Aim 2: Develop and deploy an advanced website to share CELPP results and workflows. The results of each weekly challenge across all docking methods in the challenge will be updated automatically and will provide a range of methods to analyze, compare, and download methods. The backend database will be provided by the TRDP 3 project component.

Aim 3: Analyze CELPP results to extract scientific insights and value. We will utilize the capabilities of the workflow framework to analyze the docking results, for example identifying methodological correlates of accuracy and ascertaining whether certain methods work best for certain classes or types of protein targets.

RESEARCH STRATEGY

A. Significance

The discovery of a small molecule that binds a disease-related protein with high affinity is a key step in many drug discovery projects. In the pharmaceutical industry, this step has been estimated to require over three years of work on average, at a net cost per launched drug rivaling that of clinical trials[1]. When a high-resolution structure of the targeted protein is available, structure-based computational methods may be used to accelerate the discovery of high affinity ligands. A key goal of such methods is to predict the bound conformation, or pose, of a candidate ligand. This is because an accurate method to predict the bound conformation, or pose, of a small molecule in a protein's binding pocket would allow drug designers to quickly identify chemical sites on a ligand that are disposed in space in a manner that could direct a new substituent into a targeted subsite of the binding pocket, and thus help create a new compound with greater binding affinity. In addition, accurate pose prediction, or docking, is a necessary first step in structure-based methods to predict ligand-protein affinities. More broadly, pose prediction can also provide valuable insight into biomolecular mechanisms. For example, knowing precisely how an enzyme binds its substrate is essential to understanding its catalytic mechanism. As a consequence, researchers worldwide have worked for decades to solve the so-called docking problem[2–10].

Nonetheless, computational methods for pose prediction and affinity ranking have yet to fulfill their perceived promise, as they can still be frustratingly unreliable[11]. Known challenges include estimating the energy as a function of conformation with sufficient accuracy, reliably identifying bridging waters and protonation states, and accounting for the flexibility of the protein binding[12]. As researchers have sought effective approaches to these problems, progress has come to be hampered by the perhaps surprisingly daunting difficulty of comparing the accuracy of two docking methods in a consistent and statistically meaningful fashion. This elementary but nontrivial problem has made it difficult to make and verify technical progress.

One problem is that, although new docking algorithms are frequently published with a comparison against existing methods, such comparisons are often secondary to the description of the new algorithm, and hence not fully developed. Additionally, different methods are typically tested against different sets of protein-ligand complexes, so a consistent set of comparisons is not be available. Finally, even when a study carries out careful benchmarking of multiple methods against a common dataset, the dataset often contains protein-ligand cocrystal structures that have already been published. Such retrospective studies are suboptimal, because they risk unintentional bias and because structures in the test set might have been used previously in training the docking algorithms[13]. Several initiatives have addressed these limitations through prospective or blinded, prediction challenges. In such challenges, researchers evaluate methods against a common set of test cases for which the experimental structures are withheld until after the computational predictions have been made. Prior blinded challenges include the GSK challenge[14] and CSAR[15–17]. Similarly, in recent years, the Drug Design Resource (D3R) has run blinded prediction challenges called the Grand Challenges[11,18]. These efforts have led to useful benchmarking strategies, provided insight about best practices, and sometimes yielded unexpected results regarding the effectiveness of various technical approaches. However, such episodic challenges have not been large and systematic enough to afford statistically meaningful distinctions among individual methods or to support an efficient cycle of development and evaluation that can persistently accelerate progress in the field. The present TRDP is designed to address these limitations by dramatically increasing the scale of the data flow in blinded prediction challenges of pose prediction methods, automating the collection and dissemination of data and methods, and using this framework to learn about determinants of accuracy in docking calculations (***Figure 1***).

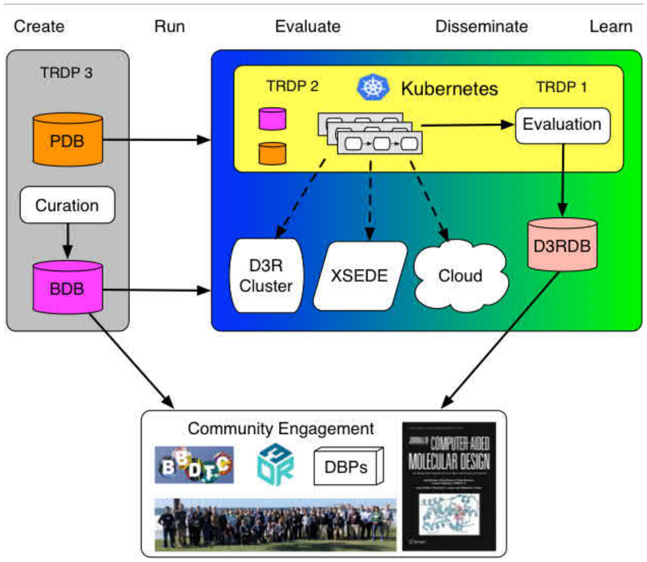


Figure 1: TRDP1 activities in the context of other TRDPs and Community Engagement.

In addition, the challenge of defining the performance of a computational method is compounded by the difficulty of reproducing complex calculations that embody many methodological choices, both explicit and implicit, and involve many operational parameters[19]. To address these issues, the community has seen a dramatic uptick in the use and availability of automated workflows that unambiguously memorialize a particular procedure and facilitate their execution and deployment[20]. The present TRDP will build on this progress, establishing procedures and technologies to capture docking methods in highly reproducible and portable containerized workflows.

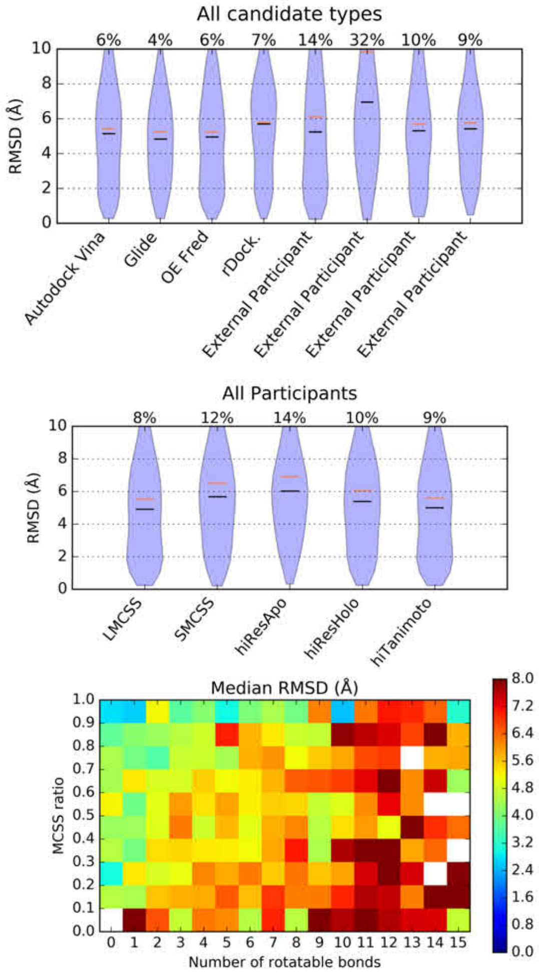


Figure 3. Analysis of preliminary CELPP results on 66 weeks of challenges. Top) Performance by method, combining predictions from all target-structure categories ("candidate types"; e.g. LMCSS, SMCSS), in terms of root-mean-square deviation (RMSD) of docked poses from crystallographic poses. Middle) Performance by target structure category, combining predictions from all participants and in-house methods. Black lines: medians; orange lines: means. Number above each dataset indicates the fraction of predictions with RMSD>10 Å. Bottom) Median RMSD as a function of number of rotatable bonds of the ligand and MCSS ratio, defined as the fraction of the heavy atoms in the docked ligand that are in its maximal common substructure with the ligand in the co-crystal structure to which is being docked. Data are taken from all methods for all calculations in LMCSS and SMCSS categories. White indicates no data.

B. Innovation

The Continuous Evaluation of Ligand Protein Predictions (CELPP) challenge[21] in this TRDP will be an innovative and indeed unique blinded prediction challenge. It is designed to overcome the limitations discussed above by directing the Protein Data Bank's (PDB)[22] ongoing stream of new structural data into a continuous docking challenge, and by utilizing new software tools to define and containerizing participants' docking workflows. Perhaps the most closely related technology is the Continuous Automated Model Evaluation (CAMEO) protein structure prediction challenge[23], which focuses on protein structure prediction rather than ligand pose prediction, and which has not so far adopted the workflow-based approach proposed here.

C. Preliminary Data

We have proven the principle of CELPP with a preliminary implementation that does not involve workflows or many other technology advances proposed below[21]. Here, we describe this implementation and key results it has provided.

Method: Each week, the PDB posts a file listing the new structures that will be released five days later. For each forthcoming structure, this list contains the PDB ID, the sequences of the proteins it contains, InChi strings of the small molecules, and the pH at which the structure was solved. Scripts developed by D3R promptly download this list and identify the forthcoming structures with protein-small molecule co-crystal structures suitable for automated docking calculations (https://github.com/drugdata/D3R). The proteins in these structures are designated as CELPP targets. Automated scripts then search the full PDB for existing structures of the CELPP targets for in the docking calculations. These structures have often been solved both without a bound ligand and also with a variety of different ligands. In order to study how the choice of target structure influences docking results, the scripts extract up to five existing structures of each CELPP target: the highest resolution apo structure (hiResApo); the highest resolution structure with any ligand bound (hiResHolo); the structure bound to the ligand with the highest Tanimoto fingerprint similarity to the ligand to be docked (hiTanimoto); the structure bound to the ligand with the largest maximum common substructure with the ligand to be docked (LMCSS); and the structure bound to the ligand with the smallest common substructure with the ligand to be docked (SMCSS). The scripts then construct the weekly CELPP data package containing, for all of the week's targets, the ligand to be docked, these suggested protein structures, and the pH. Every challenge package also includes a control challenge, which is the same every week, to aid in detection and debugging of possible anomalies. CELPP participants are responsible for downloading this data package, docking the ligands into the target structures and/or other protein structures which they draw from the PDB, and uploading their resulting predictions to a personal, password-protected web directory before the deadline several days later. Shortly after the deadline, the PDB completes its weekly structure release, which includes the crystal co-crystal structures needed to evaluate the docking predictions. CELPP scripts use these newly released structures to evaluate the submitted predictions, send the evaluation results to each participant, and add the results to running statistics available online (http://drugdesigndata.org/about/celpp).

Results: During a 66-week initial test period, this automated procedure created 1,989 docking challenges involving entirely new protein-ligand co-crystal structures drawn from the weekly PDB updates. *This number far exceeds the number of cases set by all prior blind pose-prediction challenges we are aware of, and thus strongly validates the premise of this approach.* We expect that the flow of docking challenges will continue at a similar rate, at least, during the period of this proposed project. The ligands to be docked had an average of 27 heavy atoms and 5 rotatable bonds, which is in the range of typical drug-like molecules. This flow of challenges was provided to external participants via the weekly data packages (above) and was also run on several in-house docking workflows set up by D3R as test-beds for CELPP.

Figure 3 offers several views of the performance on these data of four anonymous external methods and four in-house D3R workflows. As detailed elsewhere[21], the overall performance here is on par with that of the pose-prediction components of the recent D3R Grand Challenge 3[18], but, interestingly, not as good as in a prior blinded challenge[47]. We surmise that this is because participants in the prior study were provided with receptor structures hand-picked and prepared by human experts to accommodate the ligands to be docked. In contrast, the structures provided for CELPP docking are selected automatically and are not prepared by system experts. Similarly, Grand Challenge participants are not provided with expertly selected and prepared receptor structures.

A finer-grained analysis reveals that most methods provide rather similar levels of accuracy (**Figure 3, top**), based on median RMSD, with rDock and one External Participant performing noticeably worse. Among our in-house workflows, OE Fred and Vina yield somewhat lower RMSDs, with GLIDE and rDock trailing slightly by this metric. (Note that these workflows are not tuned for optimum results and thus may not reflect the best performance available from their respective algorithms.) The extensive data set provided by CELPP also allows quantification of important trends that have been previously noted[11,18]. First, across all methods, docking into a receptor determined with a chemically similar ligand, as determined by the maximum common substructure (LMCSS) or the fingerprint Tanimoto similarity index (hiTanimoto), more than doubles the success rate (RMSD < 2Å), relative to docking into a structure determined without a bound ligand (hiResApo). Docking into the highest resolution structure solved with any ligand (hiResHolo) and into the structure with the least similar ligand, based on maximum common substructure (SMCSS), yielded results of intermediate accuracy. The same pattern can be seen in the distributions of RMSDs in **Figure 3** (middle), and similar results are observed for each individual method. Second, docking results tend to be less accurate for ligands with more rotatable bonds, but this challenge can be overcome by docking into a protein structure determined with a highly similar ligand (**Figure 3,** bottom). In the best-case scenarios, where the ligand to be docked has 0 or 1 rotatable bonds and an MCSS ratio (see figure caption) of at least 0.8, automated docking workflows can achieve a median RMSD of around 3 Å. In more difficult cases, with MCSS ratios between 0.4 and 0.5 and 7 rotatable bonds, the median RMSD rises to about 6 Å. It should be emphasized that these quantitative analyses are possible and meaningful only because of the large number of data in this initial CELPP dataset. Development of the present TRDP will enable many more studies, which can be carried out not only by D3R team members but also by any interested researcher via the information stored in the D3RDB/BindingDB data archive (**TRDP 3**).

D. Approach

D.2 Aim1: Convert CELPP to a workflow-based challenge

1.1 Develop server environment for CELPP. We will create a compute environment (**Figure *4***) designed to support containerized docking workflows, and providing capabilities to record the entire processing history of each calculation (provenance), including all control parameters and scientific data used in the calculation and all output data, as well as performance data such as cpu time and memory usage.

Participants will submit CELPP workflows as Docker images to ensure reproducibility across a wide range of platforms. The workflow image will be fully self-contained, containing all the applications, scripts, libraries, packages, etc., necessary to run the workflow. Thus, the container brings not only the workflow of interest, but also the full system environment it requires, so researchers will be able to share and run all D3R docking workflows without installing additional software or creating a specialized run environment.

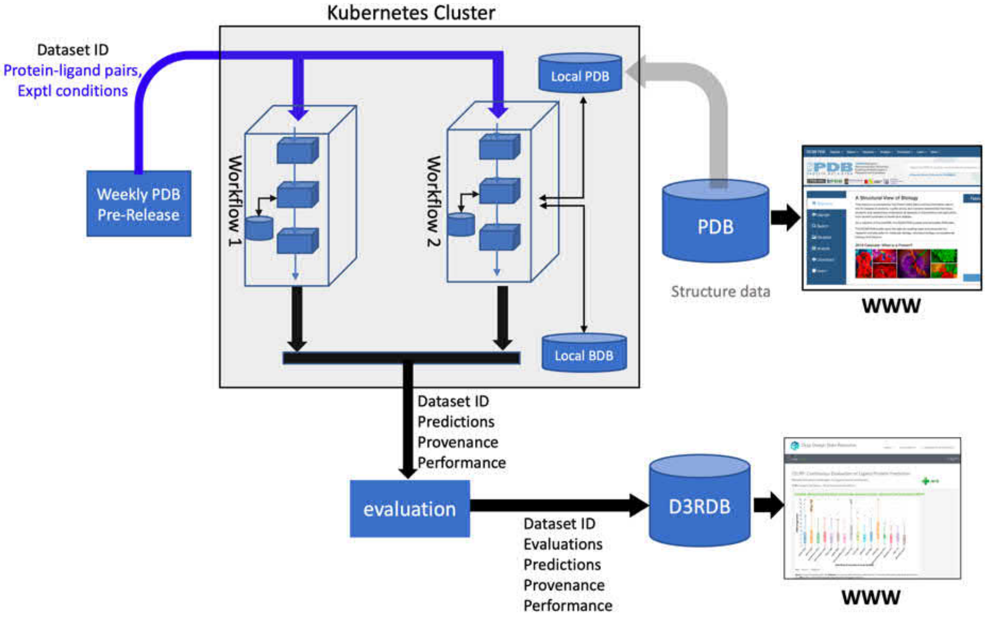


Figure 4: In TRDP 1, CELPP will take pre-release data from the PDB, feed it into containerized workflows in our Kubernetes environment, evaluate, archive, and disseminate results, evaluations, provenance, compute performance, and workflows. Two workflows are shown, each comprising three functional modules and having a persistent data store. Workflow 2 is shown as using the local PDB and BindingDB instances.

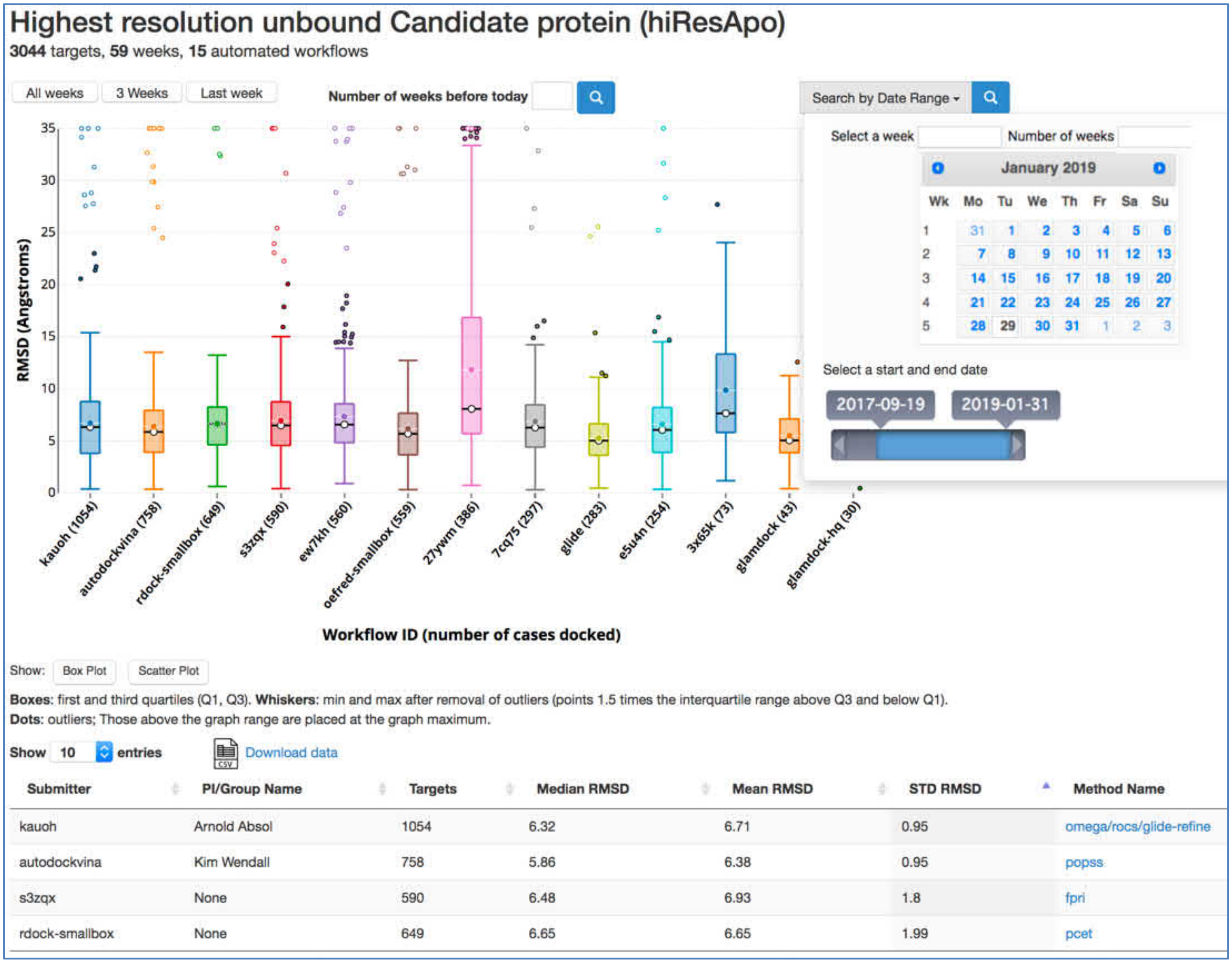
Docker is the most popular container format, and it is widely supported. Therefore, if compute demands exceed capacity on the D3R server, it will be relatively straightforward to execute participants' images on any the major cloud services, e.g., Amazon, Google, and Microsoft Azure. They can also be run on XSEDE, because it supports Singularity containers, which can be created from Docker containers. This portability also means that workflow images archived in **TRDP 3** can readily be obtained and used by other researchers.

The execution and orchestration of participant workflow containers on our server will be performed by Kubernetes, which is the open source version of Google's container management system. Kubernetes provides scheduling, fault tolerance, scalability, and elastic execution over heterogeneous resources. Containers are deployed based on a set of fine-grained metrics, e.g., required number of CPU cores or minimum amount of memory. This fine-grained resource management will allow CELPP to full control over the execution environment of the participant workflow containers. As diagrammed in **Figure 4**, the server environment will also provide up-to-date local copies of the PDB and BindingDB to workflows needing access to these data in order to set up and/or tune their calculations. Providing these locally avoids the need for any workflow to access off-site resources and thus risk unblinding a challenge or encountering network or service delays. Thus, when the workflow is run with a new dataset, a local copy of the PDB and BindingDB data is mounted in a standardized location inside the container, e.g., */data*, before the workflow executes. This allows the same workflow image to execute over new datasets released after the image was created. In addition, each workflow will be able to maintain its own local data store that will persist between challenge sets. This will enable a workflow to learn over time. (We note that specialized evaluation methods and metrics will need to be developed for workflows that adapt over time in this manner.) If we find that some workflows require access to commercial software licenses, we will maintain these licenses on the server (cost permitting) so that all workflows will operate.

1.2 Containerize D3R docking methods and deploy and test them in the CELPP server environment. We will use the four D3R docking procedures mentioned in Preliminary Data (Autodock Vina, Glide OE Fred, rDock) as test cases for containerizing pose-prediction workflows and operating them within the server environment describe above. As part of this process, we will break each method into functional modules (e.g., protein preparation, ligand preparation, docking, rescoring) that interoperate via common data specifications, and will demonstrate the construction of novel workflows by mixing and matching modules. This capability will be useful when studying the determinants of accuracy in Aim 4.

1.3 Work with DBP collaborators to develop, modularize, and install their pose-prediction workflows in the D3R server. We have already built collaborations with DBP Investigators who will work with us to deploy their own end-to-end docking workflows within the D3R CELPP framework. Community discussions will guide standardization of modules and associated data standards. Finalized Docker containers will be archived in

D3RDB/BindingDB (**TRDP 3**), along with descriptive information, information on any license requirements, etc., and will also be installed in the D3R server environment. There, they will participate in the rolling CELPP challenge, and their detailed evaluation, performance and provenance data will be archived in an ongoing manner in D3RDB/BindingDB along with the archived workflows themselves. Thus, D3RDB/BindingDB will become a unique source of operational workflows, annotated by performance and available for researchers worldwide studying computational methods and pursuing drug discovery projects.



Highest resolution unbound Candidate protein (hiResApo)

3044 targets, **59** weeks, **15** automated workflows

Wk	Mo	Tu	We	Th	Fr	Sa	Su
1	31	1	2	3	4	5	6
2	7	8	9	10	11	12	13
3	14	15	16	17	18	19	20
4	21	22	23	24	25	26	27
5	28	29	30	31	1	2	3

Boxes: first and third quartiles (Q1, Q3). **Whiskers:** min and max after removal of outliers (points 1.5 times the interquartile range above Q3 and below Q1).
Dots: outliers; Those above the graph range are placed at the graph maximum.

Submitter	PI/Group Name	Targets	Median RMSD	Mean RMSD	STD RMSD	Method Name
kauoh	Arnold Absol	1054	6.32	6.71	0.95	omega/rocs/glide-refine
autodockvina	Kim Wendall	758	5.86	6.38	0.95	popss
s3zqx	None	590	6.48	6.93	1.8	fpri
rdock-smallbox	None	649	6.65	6.65	1.99	pcet

Figure 5. Mock-up of one high-level page on the proposed CELPP website. The top left chart displays the RMSD bar plot for each of the current methods. The buttons under the header (All weeks, 3 Weeks, Last week) provide easy access to several date ranges, while the pop-up calendar on the right offers additional options. Table below the plot lists participant information and evaluation results.

D.2 Aim 2: Develop and deploy an advanced website to share CELPP results and workflows

We will develop CELPP website capabilities to provide both high-level and fine-grained analyses of the results, while also offering full downloads of data and the workflows themselves for researchers wishing to do their own analyses or to make use of selected methods in their own projects. Data will be served to the website by the integrated D3RDB/BindingDB archive (**TRDP 3**) and the PDB (via its webservices API). A mockup high-level view, **Figure 5**, shows distributions of pose RMSDs, relative to crystal structures, for a number of methods along the horizontal axis; some may be shown anonymously on the request of the developer. Website users will be able to select the target structure used (e.g., LMCSS, hiResApo) and the span of weeks for which data are to be displayed. Enhancements to the plots will include a greater number of time periods available and controls to specify time ranges; comparisons of selected methods for the largest common set of test cases they have run; evaluations for selected classes of proteins, such as kinases or aspartyl proteases; scatter plots of accuracy versus CPU or GPU time requirements; an option to present window-averaged accuracy over time for methods that learn; and an autogenerated information page for each workflow, including text documentation, citation data, information on the number of tests run and the results, and compute performance. Depending on community input, we will also consider creating an interface allowing users to submit their own pose-prediction jobs to workflows hosted on the D3R server.

D.3 Aim 3: Analyze CELPP results to extract scientific insights and value

The large number of datasets flowing through the CELPP challenge creates a new opportunity to learn about factors affecting the accuracy of pose-prediction calculations and thus advance the field. We plan to take advantage of this opportunity by assessing the domains of applicability of docking methods and evaluating the impact of various methodological component (such as protein preparation method) and parameter choices.

First, we will look for characteristics of protein targets and ligands that correlate with the performance of specific pose-prediction methods. For example, some algorithms may do better for hydrophobic sites or for specific protein families, such as serine proteases. Such statistical analyses will help practitioners choose methods suited for their specific applications and set meaningful expectations for the quality of predictions on new systems. We will also scan for docking cases where multiple methods do poorly, as these may highlight types of systems where more methods development is critically needed. Alternatively, they may help us identify cases where CELPP's automated procedures need correction. For example, one should not expect docking methods to generate accurate pose predictions if the binding site of the true co-crystal structure includes a cofactor that is absent from all structures of the target protein in the PDB. We will be alert to such technical problems with CELPP and work to correct them on an ongoing basis. Where practical, the tools used to carry out these studies will also be made available via the website (above) to support reanalysis as new methods come on line and as new challenges flow through the system.

Second, most workflows come with user-adjustable parameters, relating, for example, to the comprehensiveness of a conformational search or the weightings of various energy terms. In addition, as noted above, we will encourage developers to configure their workflows as a set of modules that can interoperate with other workflows. For example, many workflows can be configured with separate protein-preparation and ligand-preparation modules, responsible for such tasks as assigning protonation states and starting conformations. This means that, if one workflow does particularly well, we may substitute its protein- and ligand-preparation modules into other workflows and see if they lead to improvement. In effect, we will be creating new workflows and testing them in the CELPP challenge stream. This will provide new insights into the influence of not only the core docking algorithms, but also key procedural details, such as how molecules are prepared. We plan to take advantage of this modularity and the accessible parameters to seek insight into the determinants of success.

In analyzing these data, we will be mindful of and seek to account for the possible consequences of errors or uncertainties in the structural data. For example, we will work with team-member and Lead for PDB Interactions Prof. Stephen Burley to develop automated methods of down-weighting errors for ligands, or parts of ligands, whose coordinates are less well defined by the crystallographic data. We will also work with the CADD community to arrive at accepted choices for error metrics.

It is also worth noting that, if a participant's pose-prediction method were to change over time, it would become impossible to collect full statistics for a single, defined method. We will therefore distinguish between methods which are locked in for a period of time, and for which meaningful statistics therefore will be obtained, from those which are mutable, such as when CELPP is used to guide ongoing improvements in a pose-prediction method.

E. *Technology Development Integration*

TRDP3 will tightly integrate with the efforts proposed here, as the D3RDB/BindingDB archive to be developed there will store the workflows and their performance data and will serve them to the CELPP website. In addition, **TRDP2** will operate on a closely related Kubernetes framework and so we expect tight technological integration between the two efforts.

F. *Interaction with DBPs*

This **TRDP1** will be driven by a large cohort of outstanding docking methods developers, developing widely used codes (often with more than 10,000 users each) such as HADDOCK[24] (**DBP 2**), AutoDock Vina[25] (**DBP 5**), MDock[26–28] (**DBP 16**), Rhodium[29] (**DBP 1**), MolSoft[30,31] (**DBP 13**), SMINA[32,33] (**DBP 8**), MathParm[34] (**DBP 15**), and DOCK[35] (**DBP 12**). In addition, **DBPs 6, 9, 14** represent scientific teams who have partnered with us to develop and use fully vetted pose prediction workflows.

References

1. Paul, S. M. *et al.* How to improve RD productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203–214 (2010).
2. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* **161**, 269–288 (1982).
3. Goodsell, D. S. & Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics* **8**, 195–202 (1990).
4. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry* **13**, 505–524 (1992).
5. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* **15**, 488–506 (1994).
6. Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *Journal of Medicinal Chemistry* **46**, 499–511 (2003).
7. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry* **47**, 1750–1759 (2004).
8. Chen, W., Gilson, M. K., Webb, S. P. & Potter, M. J. Modeling Protein−Ligand Binding by Mining Minima. *Journal of Chemical Theory and Computation* **6**, 3540–3557 (2010).
9. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *Journal of Molecular Biology* **261**, 470–489 (1996).
10. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Computational Biology* **10**, e1003571 (2014).
11. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *Journal of Computer-Aided Molecular Design* **30**, 651–668 (2016).
12. Waszkowycz, B., Clark, D. E. & Gancia, E. Outstanding challenges in protein–ligand docking and structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 229–259 (2011).
13. Weiss, D. R., Bortolato, A., Tehan, B. & Mason, J. S. GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *Journal of Chemical Information and Modeling* **56**, 642–651 (2016).
14. Warren, G. L. *et al.* A Critical Assessment of Docking Programs and Scoring Functions. *Journal of Medicinal Chemistry* **49**, 5912–5931 (2006).
15. Damm-Ganamet, K. L., Smith, R. D., Dunbar, J. B., Stuckey, J. A. & Carlson, H. A. CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *Journal of Chemical Information and Modeling* **53**, 1853–1870 (2013).
16. Carlson, H. A. Lessons Learned over Four Benchmark Exercises from the Community Structure–Activity Resource. *Journal of Chemical Information and Modeling* **56**, 951–954 (2016).
17. Carlson, H. A. *et al.* CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *Journal of Chemical Information and Modeling* **56**, 1063–1077 (2016).
18. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *Journal of Computer-Aided Molecular Design* **0**, 0 (2019).
19. Jansen, J. M., Cornell, W., Tseng, Y. J. & Amaro, R. E. Teach–Discover–Treat (TDT): Collaborative computational drug discovery for neglected diseases. *Journal of Molecular Graphics and Modelling* **38**, 360–362 (2012).
20. Pronk, S. *et al.* Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform. *Journal of Chemical Theory and Computation* **11**, 2600–2608 (2015).
21. Wagner, J. *et al.* Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *bioRxiv* (2018).
22. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
23. Haas, J. *et al.* Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics* **86**, 387–398 (2018).
24. Kurkcuoglu, Z. *et al.* Performance of HADDOCK and a simple contact-based protein–ligand binding affinity predictor in the D3R Grand Challenge 2. *Journal of Computer-Aided Molecular Design* **32**, 175–185 (2018).

25. Forli, S. *et al.* Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols* **11**, 905–919 (2016).
26. Xu, X., Ma, Z., Duan, R. & Zou, X. Predicting protein–ligand binding modes for CELPP and GC3: workflows and insight. *J Comput Aided Mol Des* (2019).
27. Duan, R., Xu, X. & Zou, X. Lessons learned from participating in D3R 2016 Grand Challenge 2: compounds targeting the farnesoid X receptor. *J Comput Aided Mol Des* **32**, 103–111 (2018).
28. Xu, X., Yan, C. & Zou, X. Improving binding mode and binding affinity predictions of docking by ligand-based search of protein conformations: evaluation in D3R grand challenge 2015. *J Comput Aided Mol Des* **31**, 689–699 (2017).
29. Jaundoo, R. *et al.* Using a Consensus Docking Approach to Predict Adverse Drug Reactions in Combination Drug Therapies for Gulf War Illness. *International Journal of Molecular Sciences* **19**, 3355 (2018).
30. Lam, P. C.-H., Abagyan, R. & Totrov, M. Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach. *Journal of Computer-Aided Molecular Design* **32**, 187–198 (2018).
31. Lam, P. C.-H., Abagyan, R. & Totrov, M. Hybrid receptor structure/ligand-based docking and activity prediction in ICM: development and evaluation in D3R Grand Challenge 3. *Journal of Computer-Aided Molecular Design* (2018).
32. Sunseri, J., Ragoza, M., Collins, J. & Koes, D. R. A D3R prospective evaluation of machine learning for protein-ligand scoring. *J Comput Aided Mol Des* **30**, 761–771 (2016).
33. Sunseri, J., King, J. E., Francoeur, P. G. & Koes, D. R. Convolutional neural network scoring and minimization in the D3R 2017 community challenge. *J. Comput. Aided Mol. Des.* (2018). doi:10.1007/s10822-018-0133-y
34. Nguyen, D. D. *et al.* Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *arXiv:1804.10647 [q-bio]* (2018).
35. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA* (2009).

TR&D Project 2: Ligand Ranking & Affinity Predictions

SPECIFIC AIMS

A key goal of many drug discovery projects is to discover a small organic molecule that binds a targeted protein with high affinity, while also meeting requirements related, for example, to solubility, bioavailability, and pharmacokinetics. To help achieve this goal, researchers have developed computational methods of predicting, or at least ranking, protein-small molecule binding affinities. These methods range from physics-based approaches that use atomistic simulations and/or quantum chemistry, to machine-learning approaches trained on available experimental data. However, affinity prediction methods are not yet accurate enough to dramatically accelerate and reduce the costs of ligand discovery.

Part of the problem is that it is difficult to convincingly compare the reliability of various methods, to identify leading sources of error in the calculations, and to verify technical progress. This is because methods are typically tested in an unblinded fashion on datasets that differ from test to test and are too small to provide strong statistical results. The blinded D3R Grand Challenges have enabled methods to be compared without bias and on a common footing, and they have built a community around the shared goal of evaluating and advancing this technology. However, because of their slow tempo and the still modest size of their datasets, they cannot support statistically compelling conclusions about what works best and in what contexts, or provide clear feedback to developers about how to improve their methods. We propose to overcome these challenges with a novel, effectively blinded, high-throughput challenge for protein-ligand affinity prediction or ranking, preliminarily named CELPP+, after the CELPP pose-prediction challenge described in **TRDP 1**.

In brief, we will work with DBP investigators to implement their methods in automated workflows that will be hosted by the D3R team in a framework without internet access. In this setting, each workflow is effectively blinded to any affinity data that are published after the workflow has been set up, so any predictions of new data are, in effect, blinded predictions. We will then utilize BindingDB's massive flow of affinity data curated from articles and patents (**TRDP 3**) to challenge and characterize these workflows. The results will be stored in the D3RDB/BindingDB archive (**TRDP 3**), shared on a dedicated website, and mined for insights. Accordingly, the Specific Aims of this TRDP are as follows:

Aim 1 Develop workflow framework & procedures for affinity prediction workflows. Set up a compute cluster running Kubernetes to host workflows in Docker containers. Automate the uptake of newly curated data from **TRDP 3** and their application as challenge datasets in CELPP+. Automate the packaging of prediction results, run parameters, provenance data, etc., into XML files ready for ingestion by the D3RDB/BindingDB archive (**TRDP 3**) data archive. Provide workflows with access to local instances of BindingDB and PDB. Develop procedures to match the flow of challenge data fed to the speed of the workflow.

Aim 2 Populate CELPP+ framework with affinity-prediction workflows. Work with DBP investigators and two community-focused efforts, MolSSI (USA) and BioSimSpace (UK), to develop effective, standard practices for using Docker to containerize computational workflows for ligand ranking / affinity predictions. Prove principle and optimize procedures by implementing a limited set of in-house workflows, including static and adaptive (learning) methods. Move toward modularization to foster sharing and interchange of workflow components. Assist DBP investigators to use the optimized procedures.

Aim 3 Collect and share results and workflows. Develop CELPP+ website providing performance data and basic analyses for the CELPP+ workflows, downloads of detailed performance results for off-line analysis, and downloads of containerized workflows for reproducibility and use. Set up selected workflows as servers to support application-based DBPs.

Aim 4 Extract insight and value from the results of CELPP+. Use modular character of workflows to seek components that contribute to accuracy. Carry out parameters sweep of workflows with user parameters to seek optimal settings. Evaluate performance of workflows on data subsets, such as protein families or binding site polarity. Look for ongoing improvement in workflows that learn.

RESEARCH STRATEGY

A. Significance

Creating software tools capable of accurately rank-ordering a set of ligands in terms of predicted activity against a target, or of predicting the binding affinity of a ligand-protein complex outright, is a central goal in the field of computer-aided drug design[1]. Computational methods addressing this task range from physics-based approaches that use atomistic simulations[2–10] and/or quantum chemistry[11] to machine-learning approaches[12–15] trained on available experimental data. Ideally, such methods would short-cut the process of going from selection of a protein target to discovery of a potent ligand, but affinity prediction methods are not yet accurate enough to dramatically accelerate and reduce the costs of ligand discovery, and so the ligand discovery process still requires an enormous amount of trial and error, involving synthesis and assay of many candidate compounds.

A key obstacle to advancing this technology is the difficulty of rigorously comparing the reliability of various methods in order to identify leading sources of error and verify technical progress. The problem is that methods are typically tested on small, unblinded, datasets that differ from one study to another. The D3R Grand Challenges[16–18] represent an important step toward comparing methods objectively and on a common footing, and they have attracted broad participation and positive responses. Indeed, some broad lessons have been learned from these studies[16–18]. However, it is hard to extract statistically compelling, finer-grained conclusions from such challenges, because of the very limited supply of blinded, pharmaceutically relevant binding affinity data. Thus, the Grand Challenges run only about yearly and include at most several hundred test cases each, spanning one or a few drug targets.

In addition, an end-to-end affinity calculation may involve multiple software components and dozens of steps, each with its own parameters, and identifying the key determinants of accuracy in such a procedure is nontrivial. In fact, simply defining a method with enough detail to allow replication of published results can be problematic. The traditional Grand Challenges accept predictions from users, with a brief description of the method employed, but, often, not all parameters and options chosen by the participants are specified. This clearly makes the Challenge results difficult to interpret or use. In response, there is growing interest in the research community in holding challenges where participants submit not just their predictions, but also operational instantiations of their methods; see **DBP letters**. While many agree this is the necessary next step, the community is missing a standards and technical framework that would enable such approaches to be effective.

This technology development project, **TRDP 2**, addresses both limitations by creating a high-throughput, workflow-based, blinded prediction challenge for protein-ligand affinity calculations. It builds our CELPP pose-prediction challenge (**TRDP 1**), the Gilson lab's track-record of efficient, accurate curation of protein-ligand binding data from articles and patents (**TRDP 3**), and recent technologies that allow a workflow to be effectively packaged in a "container", which enables the method to be easily used on nearly any Linux machine.

B. Innovation

We propose an entirely new approach to methods assessment for computational ligand-protein affinity prediction or ranking. It has a high degree of innovation, because, simply, no one has ever done this before. In addition to providing large-scale, objective, evaluations to developers (**DBP Aim 2 investigators**) to help improve their prediction methods, this project will populate a novel archive (**TRDP 3**) of operational affinity-prediction workflows, annotated with detailed test results, which may be accessed by drug design scientists looking for the best methods for their applications (**DBP Aim 3 investigators, DBPs 6 and 14**). In later years, modularization of workflows will enable rich innovation in individual binding affinity workflow components, as well as their creative integration. Finally, the end-to-end computational workflows that are developed and enabled through this core have the potential to enable extraordinary innovation in broader applications beyond the scope of this project; e.g., **DBP 9**, an academic-industrial partnership which will utilize **TRDP 2** technology to crowd source ligand design.

C. Approach

To date, blinded affinity-prediction challenges have been intermittent and have involved data sets too small (~100 data) to support statistically strong conclusions. The problem is that few blinded data are available, despite public-spirited contributions of unpublished data by drug companies, and we see no prospect for this situation to change. The present approach overcomes this limitation by enabling the many binding data published each year to be used for effectively blinded challenges. In brief, we will collect automated, end-to-end, affinity prediction workflows from DBP collaborators, mount them in a D3R-operated framework that blocks access to data resources on the web, and challenge them on a continuous basis with the large flow of recently published protein-ligand affinity data curated from articles and patents by the existing BindingDB project, which will be integrated into this project. Because the workflows were created and isolated before publication of the data, their predictions will be effectively blinded. And because BindingDB curates tens of thousands of published affinities each year, this approach will increase the scale of affinity-prediction challenges by about two orders of magnitude, relative to the status quo. We are preliminarily calling this continuous, high-throughput affinity prediction challenge CELPP+.

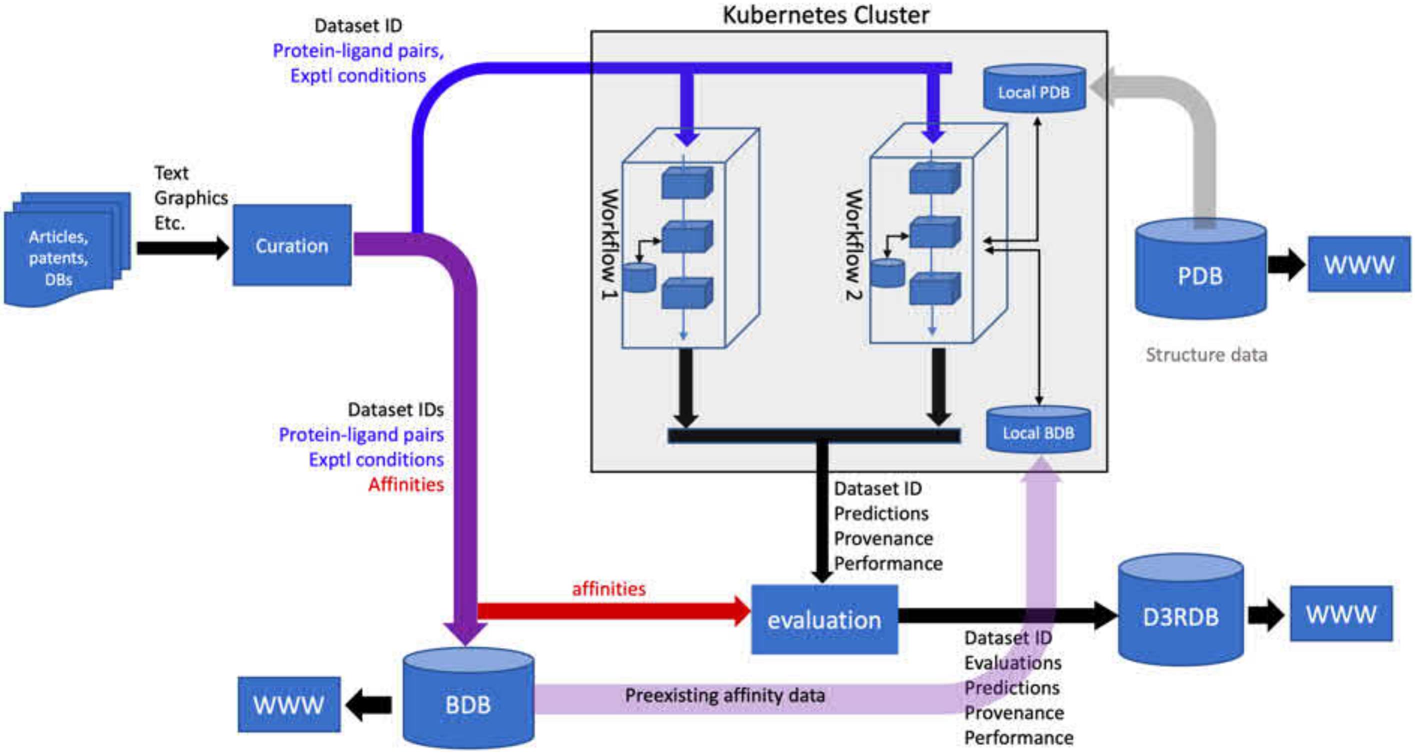


Figure 1. Data flows in CELPP+. BindingDB (BDB) staff will curate challenge datasets from articles, patents and other databases. Each dataset will have a Dataset ID, a set of protein-ligand pairs, and their affinities. The datasets (purple) flow to BindingDB for archiving and general use. In addition, the identities of the molecules (blue) flow to the Kubernetes cluster, where modular workflows (two shown here) predict the affinities. The predictions, along with data on run parameters (provenance) and compute performance, flow to the evaluation code, which compares the predicted affinities with the experimental affinities (red). The combined experimental and computational results flow to D3RDB, which is integrated with BindingDB (links not shown here). Within the Kubernetes cluster, each workflow can access a local instance of BindingDB for training and machine learning, and PDB for structure-based calculations. BindingDB, D3RDB, and PDB data are all shared publicly via the world-wide web (WWW).

C.1 Aim 1 Develop workflow framework & procedures for affinity prediction workflows

C.1.1 Technology strategy for the CELPP+ framework

We will create a container-based service which will host end-to-end workflows for protein-ligand affinity prediction, drive a flow of challenge cases through them, automatically and continuously evaluate their predictions against the associated experimental data, and pipe the results into the D3RDB/BindingDB archive (**TRDP 3)**. These proposed data flows are diagrammed in **Figure 1**. Participants will submit CELPP workflows as Docker images to ensure reproducibility across a wide range of platforms. The workflow images will be self-contained to the extent possible. That is, they will contain all the applications, scripts, libraries, packages, etc., necessary for execution. One possible exception is that some workflows may need to contact a license server for a commercial software component. We will seek to enable these applications by providing local access to licenses for widely used software, cost permitting; these can typically be situated at a fixed network address, and thus will be compatible with restricting broader internet access by the workflows, as required for “blinding”. We will collaborate with **DBPs 1-5, 7, 8, 10-13, and 15,16** to help containerize their methods with Docker. This is the most container popular format, and it will enable dissemination of methods by allowing participants' images to be executed not only on the D3R server, but also on the major cloud services, such as Amazon, Google, and Microsoft Azure. In addition, XSEDE supports Singularity containers, which can be created from Docker containers. The participant workflow images will be provided access, e.g. at a location tentatively called */data*, to local versions of BindingDB and the PDB, for use in empirical and structure-based calculations, respectively. In addition, each workflow will have the option of maintaining a local data store which will persist between calculations.

The containerized workflows will be hosted in the well-established Kubernetes framework[19], which was initially developed at Google and is now broadly supported. Kubernetes provides scheduling, fault tolerance, scalability and elastic execution over heterogeneous resources. It deploys containers based on their computational requirements, such as the required number of CPU cores or the minimum amount of memory. Kubernetes will manage the overall multi-container environment, while finer-grained scaling behavior will be managed by the workflows themselves, based on the performance requirements of the scientific executables they contain and their data management and analysis tasks. Support will be provided to record the entire processing record of each workflow, i.e., provenance, associated with parameter settings and I/O records. This data-oriented point of view, encompassing end-to-end provenance of data and workflows, enables reproducibility and effective sharing of methods. The choice of technology for the D3R shared user platform will enable us to run the resource in an effective way with reproducibility[20,21], fault tolerance, and dynamic resource coordination, building on our team's prior body of work for end to end computational and data science workflows[22–26]

To optimize the exploitation of available hardware resources, we will create a "Smart Workflow" resource selection capability, based on dynamic analysis of system states and workflow progress. This capability will predict performance of a workflow on available resources[27] and provide suggestion to users about the best infrastructure to use based on anticipated execution time and resource cost. For example, workflows that embed very computationally intensive jobs can spread the job to many heterogenous compute resources, including the Cloud or XSEDE resources.

C.1.2 Execution of a CELPP+ cycle

A CELPP+ cycle **(Figure 1**) will be initiated by the appearance of a challenge dataset created by the data curators (**TRDP 3**). When a new dataset is created, Kubernetes will schedule each workflow (**TRDP 2**) on the appropriate resources; e.g., dedicated D3R Cluster CPUs and GPUs. Detection of a new dataset may be triggered by a daemon using Linux's *inotify* to watch for file system changes in a directory, or by an SQL *trigger* when a new entry is made in the curation database. Before each workflow executes, copies of relevant data from BindingDB (**TRDP 3**) and the PDB will be mounted within the workflow container, providing fast access to the data due to locality and lack of contention with other workflows. Once the workflow completes, the results are written to persistent storage before the container is destroyed. A second Kubernetes job is then started to evaluate the results. The evaluator job compares the workflow outputs with the corresponding experimental affinities – i.e., the answers – for these input datasets and reports out its evaluations of accuracy in a data package that includes workflow ID, dataset ID, provenance, parameters, CPU/GPU time, and evaluation statistics. This package is then automatically loaded into D3RDB/BindingDB archive (**TRDP 3**), where it is linked to records of both the experimental dataset and the workflow that made the predictions.

C.1.3 Adaptations of the CELPP+ cycle

There are at least two reasons it may not be appropriate to assign every dataset to every workflow. First, if the affinity data were publicly available before the workflow was provided to D3R, then the workflow may not be blinded to this dataset. To address this, we will record the date and time when the workflow image was submitted to D3R. As the curated challenge data will include publication dates, this will make it possible to run each workflow only on data that were available after the workflow was provided to D3R. However, especially for fast workflows for which running a calculation is not burdensome, an alternative will be to run the workflow on all datasets and then address the date issues when one analyzes the evaluation data. Thus, one may choose to evaluate a workflow only with data that could not have been available when it was created, or one may choose to allow the full set of data that it was tested on. In fact, it might be interesting to see whether a workflow performs better on data that were available when it was being developed than on new data.

A second reason it may not be appropriate to assign every dataset to a workflow is that the workflow may be too slow, given available computational resources. For example, this is may apply to free energy methods based on explicit solvent molecular dynamics simulations[1]. One way to address this is to allow workflows to run asynchronously, and then define a policy for what dataset to supply next to a workflow that has fallen behind the faster workflows. Most natural, perhaps, is to continue feeding it datasets in the original order, even though this means falling further and further behind. An advantage of this approach is that it would maximize the number of datasets that all workflows have processed in common. However, we will be open to the possibility that a different prioritization of datasets may be preferable, perhaps to ensure that all workflows have been applied to a specific set of protein targets, such as kinases.

C.1.4 Preliminary Results

The proposed "Smart Workflow" Resource Selection service leverages machine learning models for creating precise performance predictions of unseen modules of a workflow[28] and we (Altintas and coworkers) have applied this technology to a compute- and data-intensive Microbiome Taxonomy and Gene Abundance workflow (MTGA)[27] . For workflow coordination, we will build on a synchronous dataflow pipeline that uses Jupyter notebooks and Kepler WebView[29] on the NSF CHASE-CI, a network of fast GPU appliances built by D3R co-Investigator Altintas and coworkers, for machine learning and storage, and managed through Kubernetes on the high-speed Pacific Research Platform (PRP) [30].

C.2 Aim 2: Populate CELPP+ framework with affinity-prediction workflows

C.2.1 Development of in-house workflows. To ensure smooth operation of the CELPP+ framework, optimize procedures, and prove the principle of this approach, we will develop several in-house, containerized affinity prediction or ranking workflows. These will also serve as examples for **DBP** developers and will generate interesting data to spark broader interest in the project, as discussed below and in the **Community Engagement** section.

Our approach will be to start by containerizing the simple docking algorithms from our Preliminary Studies (Section C.2.2) and putting these into the CELPP+ data flow. Moving forward, we will implement methods that may be more time-consuming but promise greater accuracy, including, in years 2-3 and onward, workflows that utilize the free energy methods touched on above. All workflows will be labeled (tagged with metadata), to identify particular requirements, e.g., methods that will be learning and changing their actual workflow settings and procedures on the fly, vs those workflows that will run without modification. In general, we anticipate that the containerized workflows developed in this TRD project will require a wider range of compute resources (e.g. cpu/gpu time) than those in **TRDP 1**.

C.2.2 Preliminary Studies. We created initial, end-to-end workflows based on the docking methods GlideXP[31–33] and OpenEye FRED[34]. These are already used in CELPP for pose prediction but are now applicable to ligand ranking. Our fully automated workflows require only a protein sequence, a set of ligands, and the experimental pH to run a complete, unattended calculation. Error! Reference source not found. shows preliminary results of applying these to a Grand Challenge 4 dataset[18] comprising 110 ligands with the protein target CathepsinS. These rather basic methods did not yield accurate predictions, but they illustrate complete, end-to-end automation and now only need to be containerized to be ready for CELPP+. They also provide a baseline and template to evaluate and develop more sophisticated methods.

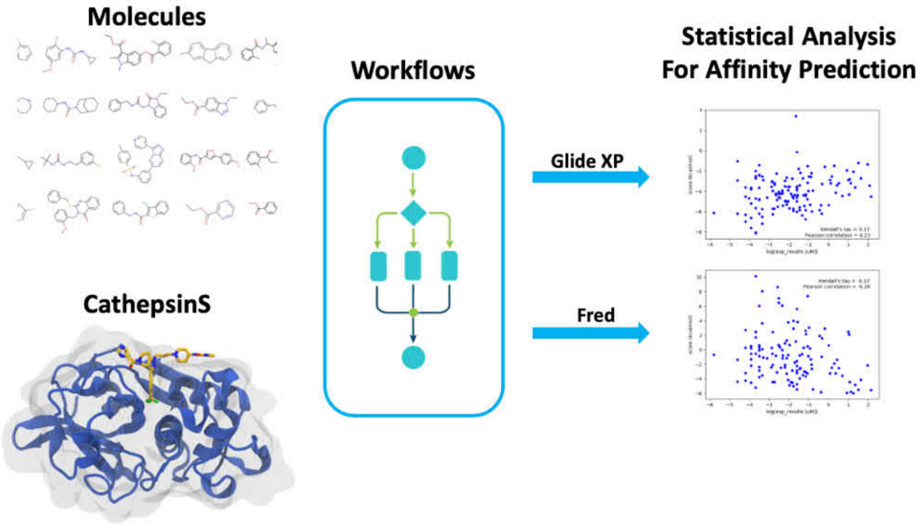


Figure 2. Two preliminary binding affinity workflows developed using the GlideXP and FRED docking programs and run on the Grand Challenge 4 Cathepsin S dataset.

C.2.3 Work with DBP Investigators and others in the community to develop their methods into workflows. The true impact of our approach will only be realized through the adoption of this approach by **DBP** investigators to create a range of additional affinity prediction or ranking workflows. Accordingly, we have laid the groundwork for a successful recruitment of developers through **13 different DBP partnerships**, each representing a leading method or software developer in the field. In addition, we have already established partnerships with leading organizations working to help the research community take advantage of state-of-the-art software development practices: the US-based, NSF-sponsored Molecular Simulation Software Institute (MolSSI, see **Letter of Support from Daniel Crawford, in Overall**); and the flagship community-oriented computational chemistry software initiative for biomolecular simulation software, BioSimSpace (see **Letter of Support from Julien Michel, in Overall**). Both of these initiatives are highly synergistic with our ongoing and proposed plans. MolSSI has committed to co-sponsorship of hackathons and workshops, development of community standards for CADD workflows, and co-hiring of a programmer who will coordinate D3R and MolSSI efforts. BioSimSpace, who we have worked with since 2017, will also collaborate on the development of global standards for workflow modules, and as part of our early adopter group for free energy method workflow development.

C.2.4 Work towards modularization of all workflows to enable module exchange and interoperability

Over time, we aim to work with the research community to break CELPP+ (and CELPP) workflows into discrete, interoperable modules, each with a well-recognized and consistent function, such as protein-preparation and ligand-preparation. Modularization will offer multiple benefits. Thus, a lab with a particularly capable protein-preparation module will be able to disseminate it to many other developers for immediate use. And the CELPP+ framework can be used to try workflows constructed with various combinations of modules, looking for best practices. Workflow modules will also support the development of corresponding training modules related to the various functionalities, as noted in **Community Engagement.**

C.3 Aim 3 Collect and share results and workflows.

C.3.1 Develop and maintain CELPP+ website, and provide avenues for archiving, sharing, and comparison of workflows.

D3R will collect workflow results in terms of accuracy and computational performance (e.g., cpu/gpu time), along with provenance data, such as the identities of the datasets used, IDs of workflows and their components, the public/anonymous flag for each workflow, adjustable run parameters, the currency dates of any databases (PDB, BindingDB), and other parameters, including information about the internal data stores of workflows which use them. Accuracy metrics will be arrived at in consultation with other researchers but will likely include Kendall's tau and Spearman's rho for ranking accuracy and root-mean-square error (RMSE) for free energy accuracy. We will continually consult with our **DBP** Investigators and the broader community to update the data that we store and evaluate. The data generated by CELPP+ will be automatically loaded into the D3RDB/BindingDB archive (**TRDP 3)** for storage, study, and dissemination.

We will develop a dedicated CELPP+ website, which will provide performance data and basic analyses for the CELPP+ workflows. The development of this website will follow the style and technical procedures established for CELPP in **TRDP 1** and will, similarly, draw on the D3RDB/BindingDB database backend (**TRDP 3**). *It will allow downloads not only of detailed performance results, but also of the containerized workflows themselves, thus establishing an entirely novel resource for drug designers to select affinity prediction methods based on their performance and computational requirements.*

C.3.2 Set up selected workflows as servers to support application-based DBPs. We aim to make selected workflows that balance speed and accuracy, and for which permissions are available, accessible as public servers. These will afford at least two means of use: 1) via a web form which allows input files and limited to be uploaded and parameters specified; 2) by a webservices API which can be accessed programmatically. This will be useful for all DBPs as well as new users beyond the initial set, and will be particularly important for **DBPs 6 and 14**, where one or more validated D3R servers will provide "second-level scoring" for crowd-sourced drug design, as detailed in the **Driving Biomedical Projects** section and the table of DBPs.

C.4 Aim 4 Extract insight and value from the results of CELPP+

C.4.1 Use CELPP+ to extracting finer-grained "lessons learned" about affinity prediction methods. D3R has sought to gain insight into the determinants of success by analyzing the results of prior Grand Challenges, as partly documented in the "Lessons Learned" page on our website. However, the small size of the Grand Challenge datasets has sharply limited what they can teach us. Because CELPP+ is expected increase the data flow by about two orders of magnitude, and because it makes each workflow available for study, it will enable statistically meaningful, fine-grained analysis of what makes some methods work better than others. For example, if one workflow performs particularly well, one may substitute its modules one by one into other workflows to learn if any in particular lead to improvement. One may also carry out parameter sweeps to systematically optimize performance. We will be interested to monitor and report back to the community on workflows that 'learn' over time, allowing new, statistically significant, insight into adaptive workflows. We will also consult with the community about what experiments would be most revealing, and, resources permitting, we can make the D3R framework available to others for experimentation.

Another approach will be to evaluate the performance of workflows on subsets of the data, categorized for example by protein family or binding site polarity. From such analyses, we expect to be able to classify easy vs. hard application classes based on prediction success. We will analyze the resulting performance data to characterize "zones of applicability" for different methods, which has been a long sought-after goal for the community and application users.

In analyzing the results, we will be mindful of uncertainties in the experimental data. We will continue to use estimates of the uncertainties (typically at least 2-3x in K_d or IC50) as a basis for bootstrapping to obtain uncertainties in computed statistics, such as Kendall's tau for ranking and Pearson's correlation coefficient for numerical data. In evaluating correlations, we will focus on the correlations within each dataset, as these will all have been obtained from one article or patent, and therefore typically derive from one lab using one method. We will also put greater weight, where possible, to Kd measurements over IC50s, although the latter will certainly be more abundant. We will also work with the CADD community to arrive at accepted error metrics.

C.4.2 Create a synergistic interplay with the Grand Challenges. Relatedly, on an annual basis, D3R will continue to run the Grand Challenges in order to allow all members of the community to test their methods. While we will continually encourage participants to develop workflows and submit them to us, as part of their submission packets, participants will still be able to submit 'predictions only' for these events. We have made this choice in order to be as inclusive as possible with regard to participant groups, recognizing that not all groups will have the desire or bandwidth to develop automated workflows of their methods (see **Community Engagement**). When certain Grand Challenge methods appear particularly promising, we will engage with the developers as **DBP** investigators to help them convert their approaches into containerized workflows that can be added to our server framework for continuous evaluation.

D. Technology Development Integration

This **TRDP 2** will use the stream of curated affinity data provided by **TRDP 3**. In addition, all of the CELPP+ data – and indeed the workflows themselves – will be archived and linked with the experimental data in the **TRDP 3** databases, which in turn will serve these data to the CELPP+ website. Furthermore, the server frameworks that support CELPP (**TRDP 1**) and CELPP+ (**TRDP 2**) will be similarly structured, so much of the work done for one will also be useful for the other. Some of the connections among the TRDPs are diagrammed in **Figure 3**.

E. Interaction with DBPs

This TRDP will interact extensively with the **DBP** Investigators who are developing methods for ligand ranking and binding affinity prediction (**DBPs 1-5, 7, 8, 10-13, and 15,16**). In addition, we will work directly with **DBP Y** investigators who will use D3R workflows to support drug design by citizen scientists, and others who will apply D3R workflows to their drug discovery and design projects (**DBPs 6 and 14**).

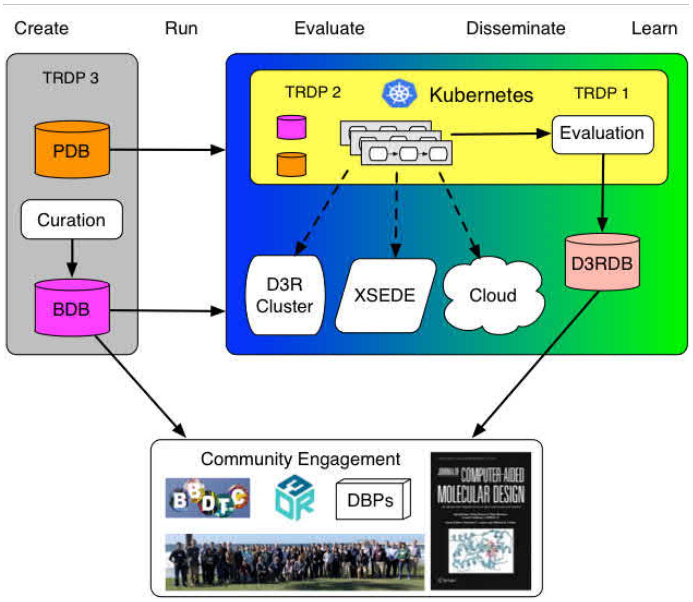


Figure 3: TRDP 2 Activities in the context of TRDP 1, TRDP 3 and Community Engagement.

References

1. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **57**, 2911–2937 (2017).
2. Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S. & Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **7**, 207–218 (2016).
3. Wang, L. *et al.* Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *J. Chem. Theory Comput.* **13**, 42–54 (2017).
4. Mobley, D. L. & Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **137**, 1–12 (2012).
5. Abel, R., Wang, L., Mobley, D. L. & Friesner, R. A. A Critical Review of Validation, Blind Testing, and Real-World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top. Med. Chem.* **17**, 1–8 (2017).
6. Jorgensen, W. L. & Ravimohan, C. Monte Carlo simulation of differences in free energies of hydration. *J. Chem. Phys.* **83**, 3050–3054 (1985).
7. Essex, J. W., Severance, D. L., Tirado-Rives, J. & Jorgensen, W. L. Monte Carlo Simulations for Proteins: Binding Affinities for Trypsin−Benzamidine Complexes via Free-Energy Perturbations. *J. Phys. Chem. B* **101**, 9663–9669 (1997).
8. Leung, C. S. F., Leung, S. S. F., Tirado-Rives, J. & Jorgensen, W. L. Methyl Effects on Protein–Ligand Binding. *J. Med. Chem.* **55**, 4489–4500 (2012).
9. Rombouts, F. J. R. *et al.* Pyrido[4,3- e][1,2,4]triazolo[4,3- a]pyrazines as Selective, Brain Penetrant Phosphodiesterase 2 (PDE2) Inhibitors. *ACS Med. Chem. Lett.* **6**, 282–286 (2015).
10. Abel, R. *et al.* Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr. Opin. Struct. Biol.* **43**, 38–44 (2017).
11. Luehr, N., Jin, A. G. B. & Martínez, T. J. Ab Initio Interactive Molecular Dynamics on Graphical Processing Units (GPUs). *J. Chem. Theory Comput.* **11**, 4536–4544 (2015).
12. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. K DEEP : Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).
13. Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L. & Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.* **53**, 592–600 (2013).
14. Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).
15. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).
16. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **30**, 651–668 (2016).
17. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* **32**, 1–20 (2018).
18. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* **0**, 0 (2019).
19. Burns, B., Grant, B., Oppenheimer, D., Brewer, E. & Wilkes, J. Borg, Omega, and Kubernetes. 24
20. Mesirov, J. P. Accessible Reproducible Research. *Science* **327**, 415–416 (2010).
21. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLOS Comput. Biol.* **9**, e1003285 (2013).
22. Kandaswamy, G., Mandal, A. & Reed, D. A. Fault Tolerance and Recovery of Scientific Workflows on Computational Grids. in *2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID)* 777–782 (2008). doi:10.1109/CCGRID.2008.79
23. Simmhan, Y. L., Plale, B. & Gannon, D. A Framework for Collecting Provenance in Data-Centric Scientific Workflows. in *2006 IEEE International Conference on Web Services (ICWS'06)* 427–436 (2006). doi:10.1109/ICWS.2006.5
24. Juve, G. *et al.* Scientific workflow applications on Amazon EC2. in *2009 5th IEEE International Conference on E-Science Workshops* 59–66 (IEEE, 2009). doi:10.1109/ESCIW.2009.5408002
25. Simmhan, Y. L., Plale, B. & Gannon, D. A Survey of Data Provenance in e-Science. *SIGMOD Rec* **34**, 31–36 (2005).

26. Juve, G. & Deelman, E. Resource Provisioning Options for Large-Scale Scientific Workflows. in *2008 IEEE Fourth International Conference on eScience* 608–613 (IEEE, 2008). doi:10.1109/eScience.2008.160
27. Singh, A., Rao, A., Purawat, S. & Altintas, I. A machine learning approach for modular workflow performance prediction. in *Proceedings of the 12th Workshop on Workflows in Support of Large-Scale Science - WORKS '17* 1–11 (ACM Press, 2017). doi:10.1145/3150994.3150998
28. Singh, A., Nguyen, M., Purawat, S., Crawl, D. & Altintas, I. Modular Resource Centric Learning for Workflow Performance Prediction. 6
29. Crawl, D., Singh, A. & Altintas, I. Kepler WebView: A Lightweight, Portable Framework for Constructing Real-time Web Interfaces of Scientific Workflows. *Procedia Comput. Sci.* **80**, 673–679 (2016).
30. Smarr, L. *et al.* The Pacific Research Platform: Making High-Speed Networking a Reality for the Scientist. in *Proceedings of the Practice and Experience on Advanced Research Computing - PEARC '18* 1–8 (ACM Press, 2018). doi:10.1145/3219104.3219108
31. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
32. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
33. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **49**, 6177–6196 (2006).
34. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **51**, 578–596 (2011).

TR&D Project 3: Data & Analytics

SPECIFIC AIMS

The overall goal of this project is to create technologies that will enable dramatic advances in two core purposes of computer-aided drug design (CADD): ligand-protein pose prediction and affinity prediction or ranking. Success in this endeavor will lower the cost and accelerate the discovery of new medications across a range of therapeutic areas. Our approach is based on a recognition that the field currently lacks effective methods to test the accuracy of CADD methods, and that such methods are needed to advance the state of the art. The problem is the shortage of unpublished data that can be used to carry out blinded – and hence objective, large-scale – and hence statistically significant – tests. We propose to overcome this problem by tapping into two large, existing flows of data in a manner that will generate effectively blinded prediction challenges two orders of magnitude larger than are possible today. The first is the flow of new PDB entries, many of which are cocrystal structures of proteins with drug-like ligands; these enable the pose-prediction challenge. The second is the flow of newly curated protein-ligand binding data generated by the BindingDB project; these will enable the affinity-prediction challenge. For pose-predictions, a blinded challenge is enabled by using the PDB's advance notice of forthcoming structures. For affinity predictions, a blinded challenge is enabled by ensuring that the predictions are made by automated workflows that are isolated from the internet before the affinity data become available.

The Data and Analytics TRDP will develop and deploy technology to collect the protein-ligand interaction data needed for the D3R challenges, and to archive and disseminate these along with data regarding the methods used in the challenges and the outcomes. This will include the shareable workflows themselves and their operating parameters, which effectively become part of the data, in order to support dissemination of these methods and reproducibility. The protein-ligand curation and archiving approach will build on the established infrastructure of the BindingDB project, which will be extended to support this project; new structural data will be linked with the Protein Data Bank.

The resulting resource will serve data to the CELPP, CELPP+ and Grand Challenge webpages discussed elsewhere in this proposal, enabling browsing, queries, selected analyses, as well as downloads of data and workflows for offline study and use by computational chemists and researchers engaged in drug design projects. The flow of curated protein-ligand data will be directed into BindingDB, which is used by medicinal chemists, computational chemists, chemical biologists, and others worldwide. In addition, we will develop an export format for the full BindingDB dataset, in order to generate the local data store in the CELPP+ workflow system which will be used by workflows that learn from available binding data.

Accordingly, the specific aims of this TRDP are as follows:

Aim 1. Collect and curate protein-ligand interaction data. For the Grand Challenges, collect hitherto unpublished protein-ligand cocrystal structures and affinities from pharma. For CELPP+, tap the much larger flow of data – tens of thousands of measurements per year – generated by BindingDB's data curation practice. Archive all incoming data in BindingDB. For CELPP, draw structural data from the weekly PDB releases.

Aim 2. Archive workflows and challenge results with associated protein-ligand interaction data. Extend the BindingDB data dictionary with tables and relationships that allow workflows to be archived along with the results of their predictions when applied to the protein-ligand datasets added to BindingDB in the course of this project. Develop software to automatically populate these data tables with the results of the ongoing CELPP+ challenges, and to support manual curation steps as needed.

Aim 3. Serve data to project websites, CELPP+ workflows and the research community. Design the local instantiation of BindingDB that will be incorporated into the CELPP+ framework and automate the provisioning of regular updates. Maintain and enhance BindingDB's web pages, APIs, and downloads to support scientific research and drug discovery worldwide.

RESEARCH STRATEGY

A. Significance

The D3R project will generate large flows of two types of valuable data: tens of thousands per year of experimental protein-ligand interaction data, curated from patents and articles and utilized to drive the automated CELPP+ challenge (**TRDP 2**); and the methods and results of the continuous CELPP and CELPP+ challenges. These large flows will be accompanied by smaller, but still valuable, flows of hitherto unpublished protein-ligand data provided by industrial DBP Investigators and other collaborators and utilized to drive the episodic, and relatively small, Grand Challenges. The methods and results of the Grand Challenges are also additional data sources that are generated. In **TRDP 3: Data & Analytics**, we will develop the technology to orchestrate these data flows, archive the data in a tightly integrated manner, and make it available for use within this project and by the broader research community. Doing so will enable transformational advances for the field of computer aided drug design (CADD), as well as aid numerous DBP investigators across academic and industrial sectors. The following two subsections discuss the significance of the experimental protein-ligand binding data and of the data on computational methods and their performance.

A.1 Protein-ligand binding data

Most drugs are small organic molecules (molecular weight below ~900 Da), which bind a disease-related protein with high affinity and inhibit or otherwise modulate its activity. In the field of chemical biology, small molecules which bind specific proteins are also used as chemical probes to delineate pathways and mechanisms. The hunt for small molecules that bind targeted proteins with high affinity generates a tremendous amount of data, notably the identities of the protein targets and the compounds made and tested, along with the potencies of the compounds against the proteins, as measured by a given assay. This information has great value, not only to the projects which generated them, but also for wider-ranging applications that become possible when a large volume of binding data can be systematically analyzed. However, the vast bulk of available protein-small molecule binding data first reach the scientific community in the form of publications: data from drug discovery projects appear in the medicinal chemistry literature and in patents, while biochemical and chemical biology studies typically appear in a different set of journals. Although publications are an essential means of scientific communication, they are incompatible with systematic search and retrieval of chemical and affinity data, or for the facile use of such data as tests for computational methods, like those organized by D3R. As a consequence, they impose a high barrier to use of the global protein-small molecule interaction dataset.

This barrier is overcome by BindingDB[1,2] (https://www.bindingdb.org) and ChEMBL[3,4] (https://www.ebi.ac.uk/chembl), the two publicly accessible database projects that engage directly in curation of protein-ligand binding data from journals and/or patents. The curation efforts of ChEMBL focus on a core set of medicinal chemistry journals, such as J. Medicinal Chemistry, Bioorganic & Medicinal Chemistry, Bioorganic & Medicinal Chemistry Letters, and the European Journal of Medicinal Chemistry. A new tranche of data is released about annually, and the ChEMBL24 release added approximately 25,000 new binding data. (ChEMBL also curates small molecule bioactivity data not definitively interpretable in terms of protein binding.) BindingDB curates another ~50,000 data per year from US patents and several chemical biology journals. BindingDB and ChEMBL exchange their data, to best serve the research community. **Thus, due to these combined curation efforts, the holdings of BindingDB are growing by ~75,000 data per year. This large flow of protein-ligand binding data will be tapped to drive CELPP+, the novel, framework for continuous evaluation of protein-ligand affinity prediction methods described in TRDP 2.**

These data will also continue to be directed into BindingDB, and BindingDB will continue to support worldwide drug discovery research part of the present project. Thus, BindingDB's website supports about 1,200 active users per week, per Google Analytics, and the data also are extensively downloaded for offline use, with over 100,000 downloads of parts or all of the dataset per year. The 2006 and 2015 updates of BindingDB in the database issue of Nucleic Acids Research[1] have been cited over 1,200 times (Google Scholar). A plurality of BindingDB web sessions originate the United States, but BindingDB usage is truly global, with users in at least four continents. Researchers at many institutions report using BindingDB for drug discovery, including University of Wisconsin - Milwaukee, St. Jude Children's Research Hospital, Yale Center for Molecular Discovery, Vertex Pharmaceuticals, Merck Research Laboratories, MRC Technology (UK), BioCryst Pharmaceuticals, Lombardi Comprehensive Cancer Center (DC), Ecole Polytechnique (France), the School of Pharmaceutical Sciences at Sun Yat-Sen University (China); and Uppsala University (Sweden). BindingDB also is used for many purposes. Survey respondents gave their main applications as drug discovery (54%), chemical biology (19%), computational methods development (12%), toxicology or environmental (6%), and systems pharmacology or systems biology (9%). Several respondents even reported clinical applications of BindingDB, such as selecting "pharmacologically rational" multidrug regimens for palliative care; and many respondents use BindingDB for teaching and learning, both in classes and in undergraduate and graduate research training programs, as well as to help with the development of texts and reference books. Finally, BindingDB is well-regarded: open, web-based surveys of hundreds of comers to the web-site have consistently yielded mean quality ratings of about 1.8 on a 1-5 scale where 1 is best.

A.2 Computational methods and their performance data

Computer-aided drug design (CADD) technologies have enormous potential to speed the discovery of new medications and lower the costs of drug discovery. When the three-dimensional structure of a targeted protein is known, two key goals of CADD are to predict the bound conformation (pose), of candidate ligands; and to predict, or at least correctly rank, the binding affinities of candidate ligands for the target[5–7]. Advances toward meeting these goals are embodied in multiple software packages, ranging from developmental academic codes to polished commercial products. Furthermore, even when a protein structure is not available, ligand-based computational methods, based on knowledge of a set of ligands known to bind the target, may be used[6,8,9]; and more recently, methods based on new machine learning approaches have been developed[10–12]. Nonetheless, these problems cannot yet be viewed as having been solved[13–17], and computational chemists engaged in drug design projects routinely face the challenge of deciding which method is best to use in a given scenario and how best to use it. Similarly, those developing new methods must put their innovations in the context of existing approaches. However, evaluations of CADD methods typically are based on experimental data known to the evaluators, and this poses a risk of unintentionally tuning methods to match the known data. In addition, different CADD methods are typically benchmarked using different datasets, so it has been difficult to compare methods on an equal footing.

In its first five years, the D3R project helped provide objective, systematic, evaluations of CADD methods by holding blinded prediction challenges called the Grand Challenges[15–17]. Although these focused researchers on the importance of rigorous evaluations and provided some insight into the effectiveness of various methods, they involve too few data to support many finer distinctions that are statistically meaningful or to allow identification of factors that drive accuracy. The present project goes far beyond the Grand Challenges by running continuous, high-throughput tests of pose-prediction and affinity assessment methods, called CELPP (**TRDP 1**) and CELPP+ (**TRDP 2**), respectively. With approximately 1,500 pose-predictions per year in CELPP[18] and potentially tens of thousands of affinity calculations or rankings per year (Section A.1), these automated challenges will generate a large volume of results. The present Data and Analytics component, **TRDP 3,** will develop and deploy the technology required to record this information, integrate it with related data, and make it available for use by this project team, by **DBPs 1-5, 7, 8, 10-13, and 15, 16** and by many others in the research community. It will thereby support a range of resources and activities. For example, this TRDP will serve data to the CELPP and CELPP+ websites in **TRDP 1** and **TRDP 2** and will also support scientific analysis of those studies. The archived records of accuracy and computational speed of various methods will also support selection of workflows suitable for deployment as Level 2 scoring engines in the citizen science project of **DBP 9**.

Finally, this TRDP will enable wide ranging investigations of the CELPP and CELPP+ results by any researcher with internet access and a computer. For example, a user may wish to analyze the performance of various docking methods on a subset of experimental data he or she deems particularly reliable, or to choose an affinity ranking method for a specific class of protein targets, such as kinases or serine proteases. For developers with workflows in the D3R system, this database will provide the means to track the performance of their successive versions over time, to identify and critically analyze outliers, and to systematically compare their results with those of others. For example, it will be possible to compare two methods by examining their results for all of the datasets they both have processed. Importantly, we will archive not only the identities of workflows and their results, but also the workflows themselves, along with full input files from each calculation, so that researchers can replicate and extend specific calculations of interest. **In summary, TRDP 3 will enable full exploitation of the challenge results generated in the course of this project.**

B. Innovation

To date, no one has previously directed a stream of protein-ligand affinity data from an ongoing curation process into a continuous prediction challenge and then routed the flow of prediction statistics into a database that integrates the experimental data with records of the computational methods and their performance. Thus, this

TRDP will create the first publicly accessible database with extensive annotation of CADD methods, including accuracy and speed, tightly coupled to the experimental data records; and to support downloading of executable workflows for further research. In addition, the curation process itself will add tens of thousands of new protein-ligand interaction data to the publicly accessible corpus each year.

C. Approach

Figure 1 provides a high-level diagram of the proposed flows of affinity and challenge result data. In brief, unpublished data provided by industry (top left) are used for Grand Challenges and are also curated and archived in BindingDB (BDB) for general use. Published data from articles and patents (left) are curated and moved into BindingDB and are also used for the CELPP+ (**TRDP 2**) challenge, which yields effectively blinded results with data published after deposition date of the workflow. The predictions, evaluations, and methodological information from the challenges are moved into the D3RDB extension of BindingDB, whose data dictionary is described below. The affinity data in BDB/D3RDB is interlinked with corresponding structural data in the PDB, as also described below. All archived data are made available via the WWW. The analogous diagram for structural data and the CELPP challenge is very similar and has been omitted for brevity. Further details of these procedures follow, by Aim.

C.1 Aim 1: Collect and curate protein-ligand interaction data

C.1.1 Collection of affinities and cocrystal structures for the Grand Challenges

The D3R project has brokered contributions of challenge-ready, unpublished data from pharmaceutical companies, and these have enabled the four Grand Challenges held to date. At this time, we have multiple datasets ready to use for future Grand Challenges, and we have current, fully executed data transfer agreements with more than 5 companies, with two additional being actively discussed. Once each challenge is complete, the affinity data are deposited in BindingDB, where they are featured on a dedicated web page (http://www.bindingdb.org/bind/ByD3R.jsp) and integrated into the full database; the cocrystal structures are curated and ingested by the PDB; and the corresponding BindingDB and PDB entries are linked. We aim to continue this process as long as the research community is interested in continuing the Grand Challenges.

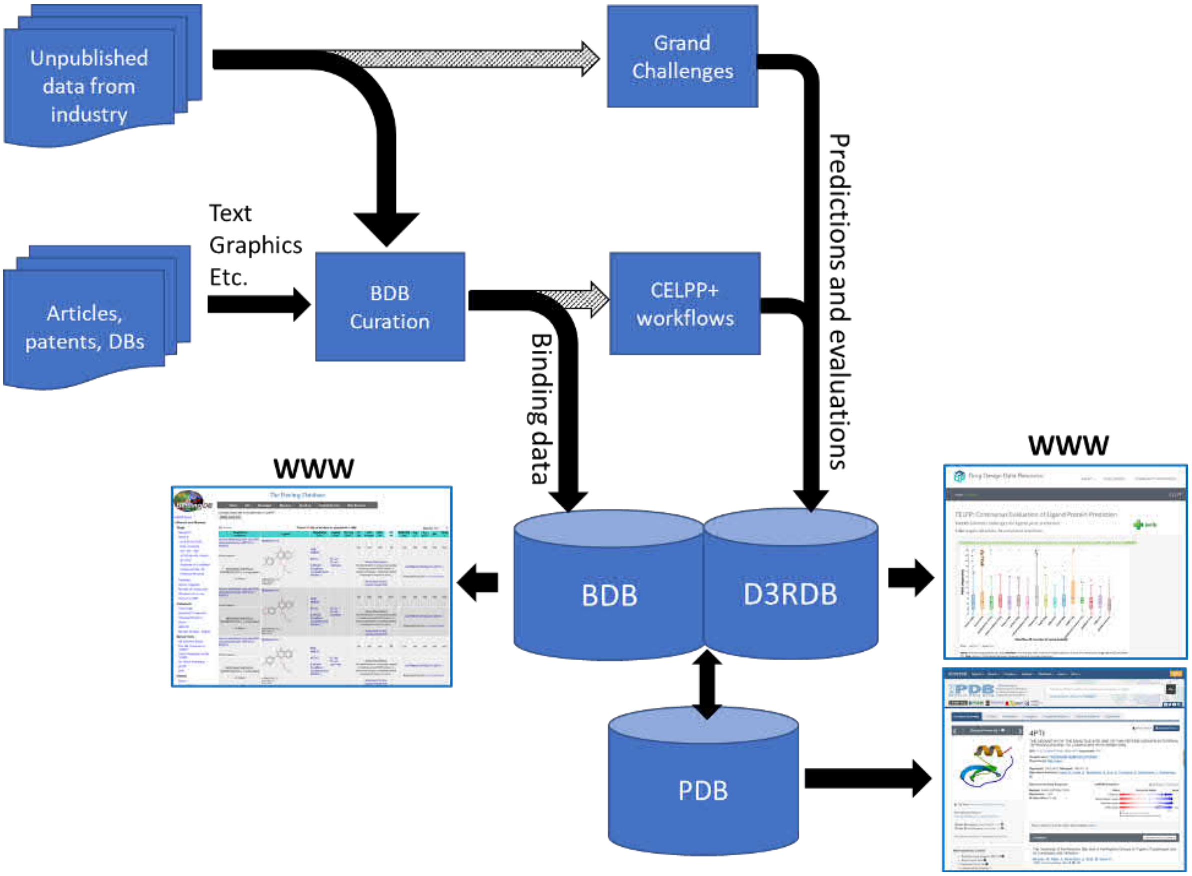


Figure 1. Data flows for affinity prediction challenges. Hatched arrows indicate challenge data “blinded” by removal of the experimental affinities and simplified by removal of needless ancillary information, such as literature citations. BindingDB (BDB) and the D3RDB of challenge methods and results are shown as fused, to indicate that D3RDB will be constructed as an integrated extension of BindingDB.

C.1.2 Collection of affinities for CELPP+

We will generate the flow of affinity data needed for CELPP+ by using BindingDB's robust, established procedures and tools to curate and collect affinity data published in patents, scientific articles. These procedures will be further accelerated by an ongoing collaboration with Prof. Leon Bergen (UCSD; see **Letter of Support**), who is using the existing BindingDB corpus to train neural networks to extract data from US Patents. We will also collect suitable data that ChEMBL has extracted from the medicinal chemistry literature. Note that CELPP+, by design, will generate effectively blinded predictions despite working on published data, as detailed in the section on **TRDP 2**. Based on experience, we anticipate curating about 50,000 data per year and collecting about 25,000 additional data from ChEMBL, where the unit is the measured affinity of one protein with one small molecule ligand. The prioritization of various classes of data to curate will be guided by our External Advisory Committee (see **Admin**) and discussions with the research community (see **Community Engagement**). Considerations may include ease of curation (patents are a particularly rich source of data); data quality and relevance (e.g. Kd vs IC50); and balance across protein (target) types.

The curation procedure may be summarized as follows. Articles and patents likely to have suitable data are identified with key-word searches. Each candidate document is reviewed by a curator and about a third are found to contain suitable data. Many US patents are associated with machine-readable representations of the compounds contained, and we take advantage of these when possible. Automated software generates a first-draft curation of the document, which is then manually corrected as needed in a web-based interface, which can be accessed by curators on- or off-site. This curation page generates an XML file containing the data and a link to the document. This file is later reopened by a second curator who checks the curation for accuracy and consults with the first curator regarding any apparent errors or omissions. Upon completion of this process, custom software extracts the data from the XML file and enters it into an off-line mirror of the public database. The new data are then processed and annotated with scripts which, for example, construct links to the PDB and UniProt, and fill out the all-by-all chemical similarity matrix used to quickly identify similar compounds on user request. Once this process is complete and the resulting data entries have passed manual review, they are moved to the public server for general use.

C.1.3 Collection of cocrystal structures for CELPP

We will continue our current approach of identifying suitable cocrystal structures in the weekly PDB releases by filtering their weekly pre-release list of structures[18], as further described in the section on **TRDP 1**. Because these data are obtained from the PDB, no additional curation or archiving work is required on our part.

C.2 Aim 2: Archive workflows and challenge results with associated protein-ligand interaction data

A stable, open, database of challenge results and associated experimental protein-ligand interaction data will provide the research community and our project team with all the information needed to derive value from the entire effort. This database will underpin our websites for the Grand Challenges, CELPP, and CELPP+ websites. Its contents will be shared under a permissive license (e.g., Creative Commons Attribution https://creativecommons.org/licenses/by/3.0/us), and it will be available for download as a database dump and in convenient simplified formats, such as tab-separated-value files of commonly used information. This novel resource will enable a wide range of analyses. For example, it will be possible to extract the data needed to compare two CELPP+ workflows, based on all protein-ligand systems to which both have been applied; to determine whether a given workflow provides more accurate results for one group of proteins (by UniProt ID) than another; to look at the performance over time of a workflow that "learns" from experience; and to compare the accuracy of automated CELPP+ or CELPP workflows with that of "by hand" Grand Challenge methods. The database will also continue to serve the worldwide users of BindingDB (see Section 2.1.1), which will come to be operated under the aegis of the present project.

The proposed database will be constructed by adding an integrated extension, containing the challenge methods and results and preliminarily named D3RDB, to the existing BindingDB schema, as indicated by the joined D3RDB and BDB databases in **Figure 1**. A simplified entity-relationship diagram of the affinities part of this extension (**Figure 2**) shows how a CELPP+ Workflow (left) generates CELPP+ Run Results, including evaluative statistics when tested on a certain date and with certain run parameters against the data in a BindingDB Entry. **Note that the actual workflow will be archived in the database along with its results and the corresponding run parameters. This innovative tactic will enable reproducibility and facile dissemination of functioning CADD tools.** We will integrate the Grand Challenge methods and results with BindingDB in a similar manner. As sketched in **Figure 2** (top right) a Grand Challenge can comprise multiple Grand Challenge

Datasets, each with one protein and multiple ligands, whose affinities are archived in BindingDB Entries. When a research group submits predictions for a Grand Challenge Dataset, the submission is assigned a Submission ID and the predictions are compared with the experimental data to yield evaluative statistics, as captured in the GC Prediction Results table. Database extensions for CELPP and the pose-prediction components of the Grand Challenges will be analogous, integrating with BindingDB in cases where affinity data are available, and with the PDB for structural data.

Data integration with the PDB will be strengthened by collaborating to make maximal use of the PDBx/mmCIF data dictionary and foreign keys within the PDB. The PDBx/mmCIF dictionary and format[19] are machine-readable and fully extensible; they assure that macromolecular structure data can be effectively represented and broadly distributed. The data dictionary is regularly updated to reflect evolution of the underlying science and technology. The current version of the dictionary (5.303), which includes >4,300 data items, is organized as 553 interlinked categories, clustered into 41 groups that describe atomic coordinates, polymer and small molecule chemistry, details of sample production, experimental setup, data collection and processing, and cross-references to other biological databases. To ensure consistency, more than 25% of the data items are subject to controlled vocabularies, which are periodically updated[20] . The rich information content of PDBx/mmCIF will support explicit mappings between D3RDB/BindingDB and the PDB and allow us to obtain structural data efficiently via PDB webservices, rather than having to store copies of the structural data, thus wasting resources and increasing the risk of denormalization.

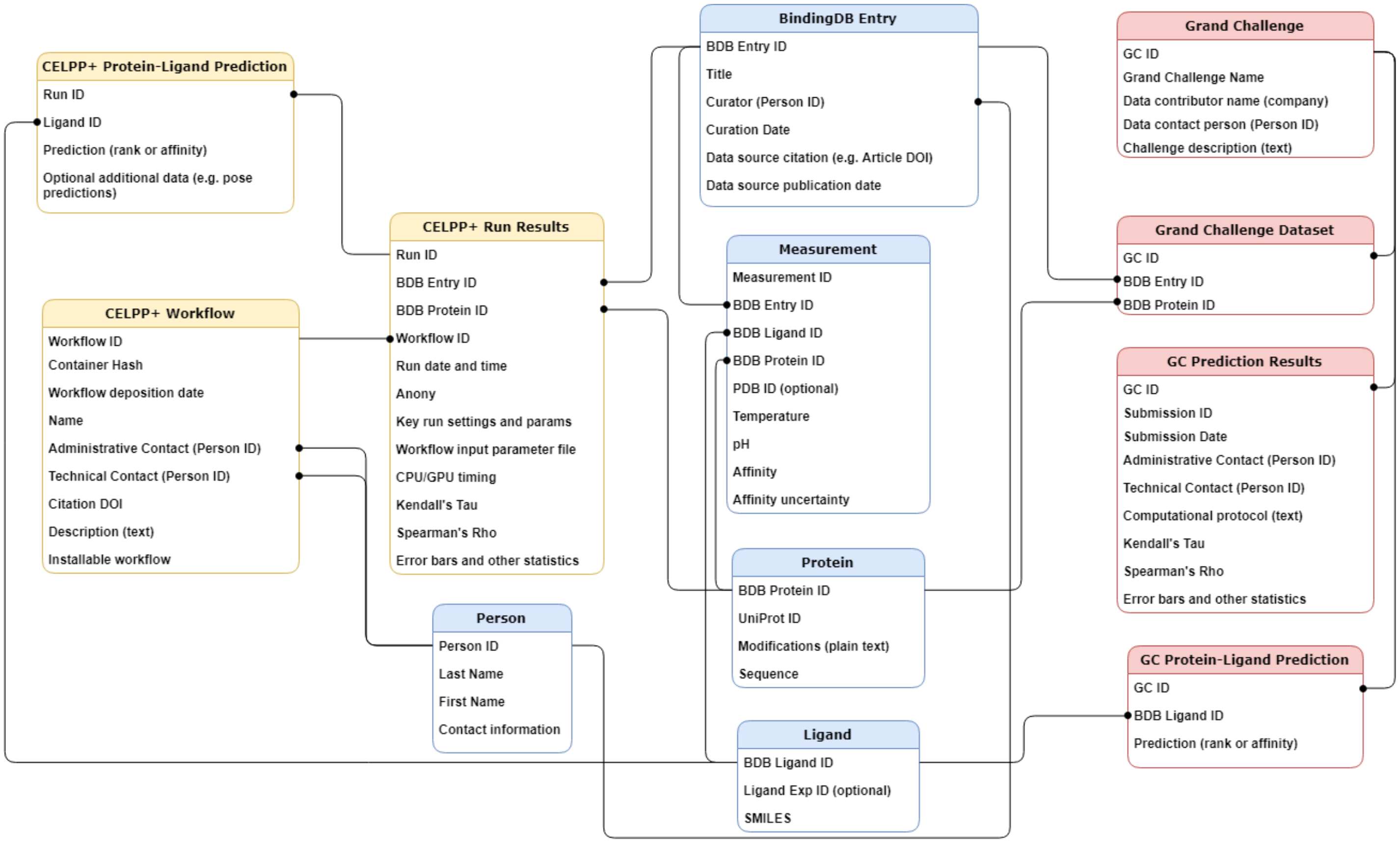


Figure 2. Simplified entity-relationship diagram of the affinities components of the proposed D3RDB/BindingDB database. Additional extensions will include pose-prediction methods and results. Blue: existing BindingDB entities; Yellow: added CELPP+ entities; Pink: added Grand Challenge entities; GC: Grand Challenge; BDB: BindingDB.

Broader data exchange will be enabled by using the University of California EZID facility, which is already used for BindingDB Entries, to assign DOIs to key D3RDB digital objects, such as workflows and challenge result sets. These digital objects will be annotated with informative metadata and additional DOIs (e.g., for articles) and database IDs (e.g., for proteins in UniProt) in order to enable cross-indexing. Note that DOIs assigned by EZID automatically appear in DataCite, and the data-literature interlinking service Scholexplorer imports DataCite DOIs and their associated metadata, thus enabling article-data links.

Procedures will be established to efficiently populate the D3RDB database components. Because Grand Challenges occur only about once a year and typically involve only a few hundred data, these data can be handled only semi-automatically. Here, user-initiated scripts will package submissions and results into XML files that can be viewed and edited, along with the corresponding experimental data, with an extension of the existing BindingDB data curation tools (Section 0). In contrast, the high throughput of CELPP and CELPP+ require that fully automated procedures – which nonetheless allow for human oversight and intervention – are required. Thus, the systems that automatically run CELPP and CELPP+ (**TRDP 1** and **TRDP 2**, respectively) will be outfitted with scripts to generate XML files ready for loading into the off-line curation mirror of the database. Especially during an initial development phase, these data will be reviewed manually for possible errors, using extensions of the existing curation tools. Apparent errors will be corrected before the data are migrated to the public database server and will also be used to guide corrections to the automated processes.

The database itself will be implemented with Oracle software, which is fast, robust, and free for academic use. However, database dumps compatible with the open-source MySQL database will be made freely available, along with Oracle dumps. Following existing practice at the PDB and BindingDB, D3RDB will use ChemAxon tools, which are available at no cost for academic projects, to manage chemically aware indexing and searching. For operational and data stability, the database will be hosted on a main server situated at the San Diego Supercomputer Center and mirrored on at least one additional server at a different location.

C.3 Aim 3: Serve data to project websites, CELPP+ workflows and the research community

Data in the D3R/BindingDB archive will be shared with a varied user-base by a variety of routes. Protein-ligand affinity data will be provided to the research community via existing BindingDB webpages, webservices, and downloads. New APIs and middleware will be developed to support specialized browsing, query, and downloads via this project's existing Grand Challenge website, and via the CELPP and CELPP+ websites to be advanced in **TRDP 1** and **TRDP 2**, respectively. Finally, we will design the "local BindingDB" and "local PDB" data caches required to support workflows in the CELPP and CELPP+ frameworks and develop the code to generate these files as needed. As noted in the other TRDPs, these compact local versions of the two databases will provide the workflows with low-latency, high-bandwidth access to the binding and structural data needed for their calculations, while avoiding the need to provide workflows with internet access, which would open the possibility of unblinding their calculations. These caches will address, for example, the facts that some affinity estimators in CELPP+ will need to train themselves on available data, and that pose prediction methods in CELPP need access to protein structures suitable for their docking calculations. The local data caches will be time-stamped, and the time-stamps will be recorded as part of the provenance of each workflow calculation and ultimately recorded in the D3RDB archive as part of the run information.

D. Technology Development Integration

This Data and Analytics TRDP 3 integrates tightly with **TRDP 1** (pose prediction) and **TRDP 2** (ligand ranking / affinity prediction). In addition to generating the flow of protein-ligand binding data that is the life blood of **TRDP 2**, the **TRDP 3** data archive will automatically capture all of the **TRDP 1** and **TRDP 2** inputs, outputs, and workflows, make them available and analyzable, and serve them to the respective CELPP and CELPP+ websites.

E. Interaction with DBPs

The work of this TRDP is strongly driven by the needs of all **DBP** investigators. Those developing affinity prediction methods require the data stream to be provided by **TRDP 3**, and both pose-prediction and affinity-prediction developers will benefit from a smart database that will archive their prediction data, their workflows and associated metadata and performance data. For example, these data will enable studies of the determinants of accuracy that will be used to drive methodologic improvements. At the same time **DBPs 6 and 14**, who are applying pose-prediction and affinity-ranking methods, will benefit from access to downloadable, operational workflows, and the capability to select workflows for given projects based on accuracy (global or for systems of interest) and/or computing (e.g., CPU/GPU) needs, as some users will have more hardware resources than others. Along similar lines, **DBP 9**, which seeks to enable crowd-sourced drug design, will benefit because our collaborators will be able to use the data in this archive to choose the methods/workflows most suitable for 2nd-level scoring. Finally, our **DBPs 1-5, 7, 8, 10-13, and 15,16** (and others) will benefit from the intelligent connection to other existing data efforts, such as NIH FAIR[21], the PDB, and BindingDB, for example.

References

1. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucl Acid Res* **35**, D198–D201 (2007).
2. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
3. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
4. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).
5. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
6. Prathipati, P., Dixit, A. & Saxena, A. Computer-aided drug design: Integration of structure-based and ligand-based approaches in drug design. *Curr Comput-Aid Drug Des* **3**, 133–148 (2007).
7. van Montfort, R. L. M. & Workman, P. Structure-based drug design: aiming for a perfect fit. *Essays Biochem.* **61**, 431–437 (2017).
8. Golbraikh, A. & Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J Comput Aid Mol Des* **16**, 357–369 (2002).
9. Hawkins, D. M., Basak, S. C. & Shi, X. QSAR with few compounds and many features. *J Chem Inf Comput Sci* **41**, 663–670 (2001).
10. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks - Journal of Chemical Information and Modeling (ACS Publications). Available at: https://pubs.acs.org/doi/abs/10.1021%2Facs.jcim.7b00650. (Accessed: 8th February 2019)
11. Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L. & Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.* **53**, 592–600 (2013).
12. Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinforma. Oxf. Engl.* **26**, 1169–1175 (2010).
13. Warren, G. L. *et al.* A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **49**, 5912–5931 (2006).
14. Dunbar, J. B. *et al.* CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *J. Chem. Inf. Model.* **51**, 2036–2046 (2011).
15. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **30**, 651–668 (2016).
16. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* 1–20 (2017). doi:10.1007/s10822-017-0088-4
17. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* (2019). doi:10.1007/s10822-018-0180-4
18. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking | bioRxiv. Available at: https://www.biorxiv.org/content/10.1101/469940v1. (Accessed: 8th February 2019)
19. Fitzgerald, P. M. D., Berman, H. M., Bourne, P. E., McMahon, B. & Westbrook, K. W. andJ. The mmCIF Dictionary: Community Review and Final Approval. *Int. Union Crystallogr. Seventeenth Congr. Gen. Assem. Acta Cryst* **A52 Suppl.**, (1996).
20. Young, J. Y. *et al.* Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database* **2018**, (2018).
21. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, (2016).

COMMUNITY ENGAGEMENT

SPECIFIC AIMS

Community engagement is of central importance to the Drug Design Data Resource (D3R) project, because it requires recruiting the developers of methods for computer-aided drug design (CADD) as participants in large-scale blinded prediction challenges. The many letters of commitment already in hand document strong community interest in and support for this enterprise, but it is essential to continue recruiting new participants in order to maximize the impact of this work on the state of the art in CADD technologies. Perhaps the strongest argument for participation is that the data generated by the new continuous challenges will enable developers not only to test, but also to improve the accuracy of their methods in ways never before possible. It will also be essential to help our DBP investigators overcome the technical and conceptual barriers of creating the hardened, containerized workflows needed for them to function in the high-throughput automated challenges to be developed here. We will address this need by collaboratively developing accepted methods and standards, and then holding hackathons and training sessions for current and prospective DBP investigators. Finally, the impact of this project will be maximized by our use of multiple channels to disseminate results and methods and to obtain valuable feedback and guidance that will help us best serve the research community.

Accordingly, our Specific Aims for community engagement are to:

Aim 1. Engage the CADD community scientifically through blinded prediction challenges. We will continue the existing Grand Challenges as highly visible events that focus community attention and provide an on-ramp for new DBP investigators wishing to use the new continuous, high-throughput challenge servers. We will also ramp up the continuous CELPP and CELPP+ challenges, recruiting new DBPs to take advantage of these novel technologies.

Aim 2. Expand technology adoption through workshops, hackathons, and online training modules. We will continue the successful annual D3R workshops, add training events and on-line modules to share CADD expertise with researchers wishing to learn best practices, and hold hackathons to help DBP investigators adapt their non-automated their methods for the high-throughput CELPP and CELPP+ servers.

Aim 3. Use electronic media to share information and interact with participants and the broader community of interest. We will build out the D3R web-presence to include the new challenges, providing access to data and workflows, and make thoughtful use of additional tools like email and Twitter to maintain contact with interested researchers while reaching out to a broader audience.

Aim 3. Disseminate and publish data, technologies, and results. We will use the D3RDB/BindingDB database, the PDB, and the D3R website to disseminate curated protein-ligand binding data, operational workflows annotated with their predictions of the experimental affinities and of ligand poses, and newly released protein-ligand co-crystal structures, while broader information about the project and its results will be shared via the peer-reviewed scientific literature and at scientific meetings.

RESEARCH STRATEGY

A. Significance and Background

Community engagement is at the heart of D3R. Every aspect of our activities, from community challenges to data acquisition and curation activities to technology development, training, and administration derives entirely from the needs and input of the computer aided drug design (CADD) community (**Figure 1**).

The CADD community that D3R serves comprises several different stakeholder groups. First and foremost, we serve the CADD method developers; these are the teams, primarily academic but also some industrial, that develop new methods for pose prediction and ligand ranking or binding affinity prediction. The methods developed by these teams have enormous use in the biomedical research enterprise, with a number of our method developers (Driving Biomedical Project (DBP) investigators) having >12,000 registered users worldwide; examples, include AutoDock Vina[1], HADDOCK[2], and DOCK[3]). D3R also serves the industrial CADD community, expert users within pharmaceutical and biotech companies (see **Letters of Support in Overall from Tara Mirzadegan (Janssen), Richard Lewis (Novartis)**, **Georgia McGaughey (Vertex),** and in **Admin: Pat Walters (Relay Therapeutics), Hanneke Jansen (Novartis), Martin Stahl (Roche);** and in DBP: **Gruson & Steuer on behalf of Andreas Bergner (Boehringer Ingelheim)**. These individuals are generally established users of CADD technologies who have a strong practical interest in knowing the state of the art, understanding the best ways to use each technology, and seeing the power of CADD methods increase. They have a powerful biomedical impact by using CADD tools daily to help create new and improved medications. The pharmaceutical and biotech industries also contribute to D3R activities by providing hitherto unpublished datasets suitable for use in our blinded Grand Challenges. Researchers using CADD methods for drug design in the academic setting are also part of our community. Their focus is similar to that of their industrial colleagues, except that they may have a stronger interest in open-source methods. The mutual engagement of the entire community brings substantial impact to the research enterprise and overall public health.

Figure 1. Some of the CADD community we serve, at our 2016 in-person workshop (top), and again in 2018 (bottom).

B. Innovation

D3R is committed to innovating in all aspects of this project, including community engagement. In fact, the rolling, or continuous, prediction challenges at the heart of this project -- i.e., CELPP and CELPP+ -- are about engaging the community in wholly novel ways to test and improve their CADD methods. We will also innovate by expanding the circle of participating developers and users; working with the community-oriented, NSF-funded MolSSI initiative (see **Letter of Support in Overall from Daniel Crawford**) and the UK-based BioSimSpace project (see **Letter of Support in Overall from Julien Michel**) to develop community-driven approaches to CADD workflow design, modularization, and containerization; holding a new series of hackathons (below) to help developers employ these new technologies; and taking advantage of innovative new frameworks for sharing of training modules, such as the Biomedical Big Data Training Collaborative (below).

C. Approach

As a central hub for CADD developers worldwide, D3R maintains an exceptionally high level of visibility, and we have already established a solid track record in community engagement. Our approach to community engagement has been multifaceted. In addition to organizing and running the Grand Challenges[4–6], we also initiated the CELPP[7] continuous challenge, hosted workshops, published in peer-reviewed journals, and participated in high profile conferences. These and other community-engagement activities will be continued and expanded by the proposed BTRR.

We also will also continue to use D3R to engage the community in intensive discussions around important dimensions of the field. For example, at our 2018 workshop, after a presentation from our external evaluator Pat Walters (Relay Therapeutics, see **Letter in Admin**) about his analysis of the whole series of Grand Challenges to date, workshop participants agreed to develop a working group to develop generally accepted evaluation metrics. Thus, there is substantial opportunity to leverage community engagement activities to enrich our resource activities. This includes opportunities to develop CADD standards, spanning all aspects of the end-to-end process, including data i/o standards, process/ methods standards, evaluation standards, and publication standards.

Formal and informal interactions with the D3R External Advisory Board will also continue to provide project guidance and feedback from various relevant sectors of the research community, including industrial users of CADD software, software developers and technology experts, and researchers with experience holding blinded prediction challenges in the protein-structure prediction space.

The following subsections provide further detail on our Aims for community engagement, while Sections D and E address collaboration and service, the technology development arc, the expected life of the center, and sustainability.

C.1 Engage the community scientifically through blinded prediction challenges

Blinded prediction challenges are central to the concept of this project. Indeed, the main purpose of the technologies to be developed is to enable a new class and scale of challenges designed to take the field to a new level. However, the more conventional Grand Challenges, described below, have laid the groundwork for the present proposal by developing procedures, generating initial results, and building a strong community of developers committed to engaging in and benefitting from blinded challenges in general. Because they do not require the added work of creating an end-to-end automated workflow, like the continuous CELPP and CELPP+ challenges, they also have a lower barrier to entry. This makes them more broadly appealing and positions them as a valuable "on-ramp" for developers who will ultimately join as DBP participants in the high-throughput CELPP and CELPP+ exercises to be developed here. Therefore, we plan to continue the Grand Challenges for at least the first several years of this project, with the understanding that they are likely to become less relevant as the new continuous challenges come on line and gain broad acceptance. Here we provide additional information about both the Grand Challenges and the continuous, or rolling, CELPP and CELPP+ challenges.

C.1.1 Grand Challenges

In the roughly annual Grand Challenges, CADD method developers and users test their ability to predict the results of hitherto unpublished experiments. These data are typically provided by pharmaceutical companies, which have large databases of unpublished information, some of which can be freed for use by D3R as the corresponding drug discovery projects end. The data that we look for are mainly X-ray crystal structures of druglike molecules in complex with a target protein, with resolution better than ~2.7 Å, along with the measured affinities of these compounds for the target, along with a larger set of compounds for which co-crystal structures are not available. A data set usually has tens of co-crystal structures and one to several hundred affinities (**Figure 2**). The co-crystal structures allow us to challenge pose prediction (or docking) methods in terms of their ability to accurately predict a ligand's bound pose, while the affinity data challenge methods to compute either numerical (absolute or relative) binding affinities, or at least the correct ranking of ligand affinities. Before any structural data are used for a challenge, they are reviewed and often re-refined in the lab of co-investigator Prof. Stephen Burley (Rutgers U; Director of the RSCB PDB), in order to maximize the quality of the structures used. This process also identifies ligands or parts of ligands that may not be well-resolved, information which is useful when it comes to evaluating the predictions. The affinity data ideally take the form of K_d's, but due to data availability, we often must work with IC50 data, which, if needed, can often be converted to K_d's via the Cheng-Prusoff relation[8]. It is also worth noting that IC50 data alone are often sufficient, because many researchers are primarily interested in correctly rank ordering the compounds by affinity, since in real world applications, ranking alone suffices to allow medicinal chemists to prioritize candidate ligands for synthesis and testing[9–11]. That said, a subset of the community that we serve, including many in industry, are also interested in

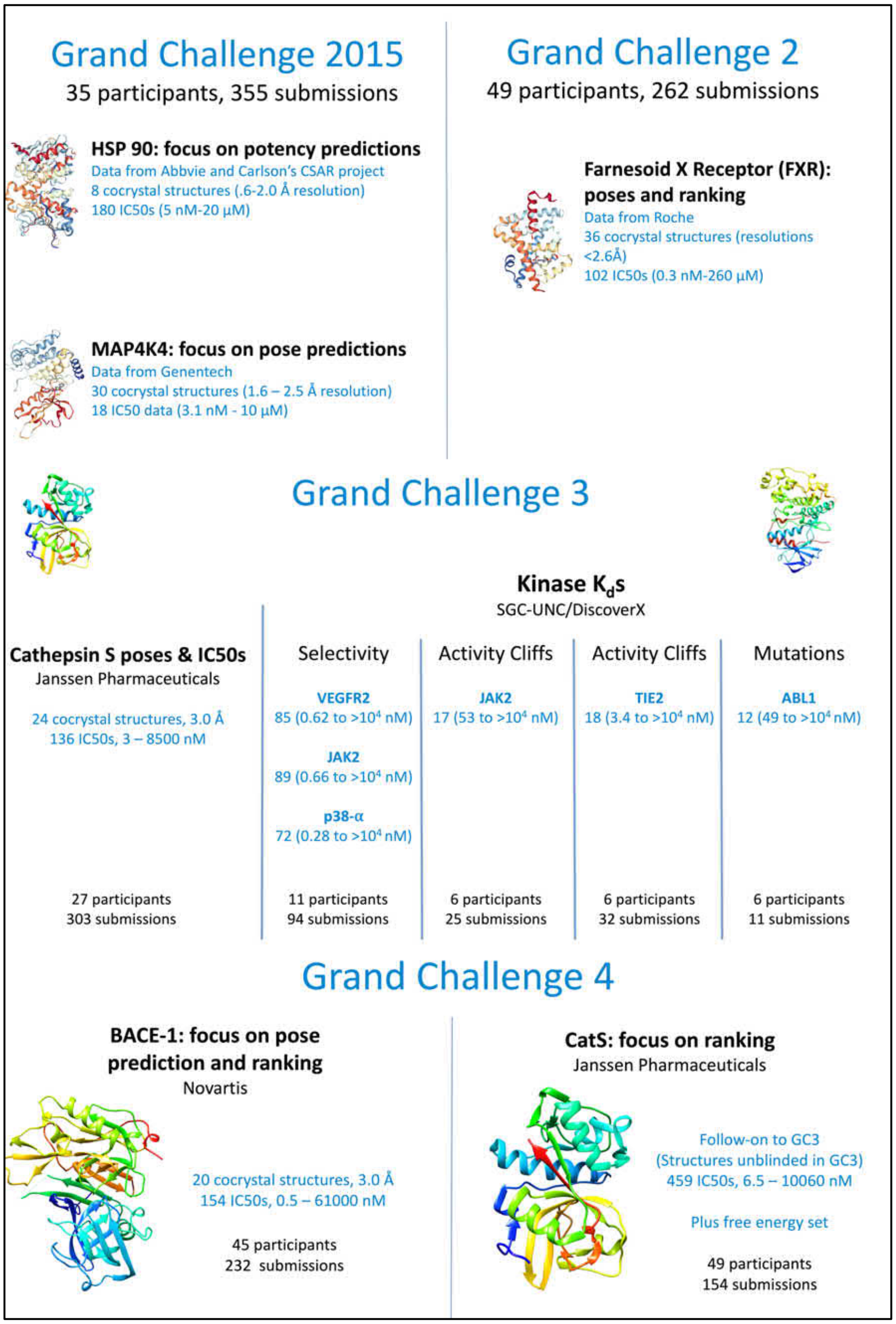


Figure 2. Summary of Grand Challenges to date; Grand Challenge 4 is currently in progress.

more detailed free energy methods, which may ultimately provide the most accurate and universally applicable approach to the questions most relevant to drug discovery and design[12–16].

At the outset of a Grand Challenge, participants are first given the identities of the target proteins and ligands, along with experimental conditions and background information on the system, such as key articles and notes on whether there is a highly mobile loop, for example. The Challenge then unfolds in several stages over about three months. In Stage 1a, participants are invited to predict the poses of all the ligands for which co-crystal structures are available, and also to predict or at least rank all the available affinities. Because participants do not have the crystal structures of the proteins solved with the ligands, the pose-prediction component of Stage 1a is considered a cross-docking challenge. In Stage 1b, participants are then provided with the crystal structures of the protein-ligand complexes, but with the ligands removed, and are invited again to predict the poses. Because they now know the structure of the protein solved with each ligand, this is considered a self-docking challenge. Finally, in Stage 2, the full co-crystal structures are provided, and participants are again invited to predict or rank all of the affinities. As each stage finishes, the D3R team evaluates all predictions and shares these evaluations with the respective participants, so they can help us correct any possible errors. Note that our evaluations of the affinity predictions or rankings include bootstrap resampling based on the known or estimated uncertainty of the experimental data. In addition, all D3R evaluations are repeated and extended by external evaluators (recently the industrial computational chemists Patrick Walters, Neysa Nevins and Millard Lambert) for quality control and further insight. The results are all then posted to the D3R website, along with the detailed experimental data. Each participant may choose to be identified or to have their data posted anonymously. As detailed below, each Grand Challenge is followed by a workshop – either in person or virtual – and is associated with a Special Issue of the Journal of Computer-Aided Drug Design (**Figure 3**), in cooperation with Editor-in-Chief Dr. Terry Stouch (see **Letter from Terry Stouch in Overall**). The special issue contains an overview of the challenge written by D3R, the data contributors and the external evaluators, along with focused articles from most of the participants.

JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN
JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN
JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN
GC 2015
14 articles, 2016
GC2
23 articles, 2017
GC3
~15 articles, 2019

Figure 3. D3R's annual special issues of Grand Challenge reports, organized in JCAMD, are highly cited and provide the community with an effective mechanism to disseminate their findings.

Since its inception in 2014, the existing D3R project has completed three Grand Challenges[4–6] and is near the end of Grand Challenge 4. The level of participation has grown in this period from about 40 participants submitting 180 prediction sets to 55 participants in 2018 submitting about 400 prediction sets. (Note that one participant may submit predictions from more than one method.) Most participants are in the USA or Europe, but others are located in Asia, Africa and South America as shown for Grand Challenge 4 in **Figure 4.** The companies that contribute unpublished data for these challenges are also span multiple countries, including the USA, Switzerland, and Austria. This global interest in and support for D3R reflects the uniqueness and value if its role in the computer-aided drug design community.

As detailed elsewhere, the continuous challenges at the center of this project are expected to boost the number of available challenge cases (poses and affinities) each year by about two orders of magnitude, relative to the Grand Challenges. This increase will provide dramatically higher resolution in comparing methods. However, the Grand Challenges bring together the community in a special and highly effective manner, as emphasized to us by SAB member John Moult. In addition, CELPP and CELPP+ will not come on line immediately. Therefore, we plan to continue the Grand Challenges on an approximately annual basis at least for the first few years of this project. Later, depending on feedback from participants and the External Advisory Committee, it may prove appropriate to sunset the Grand Challenges and focus exclusively on the high-volume continuous challenges.

Importantly, we expect that our industry partners will be able to continue providing the required datasets, as reflected in many Letters. We have active Data Transfer Agreements with prior data contributors, including Novartis, Glaxo Smith Kline, Genentech, Roche, and Janssen, and, in addition, we are working on a new DTA with Boehringer Ingelheim as well as a tripartite DTA with a different unit in Novartis and the Dana Farber Cancer Institute.

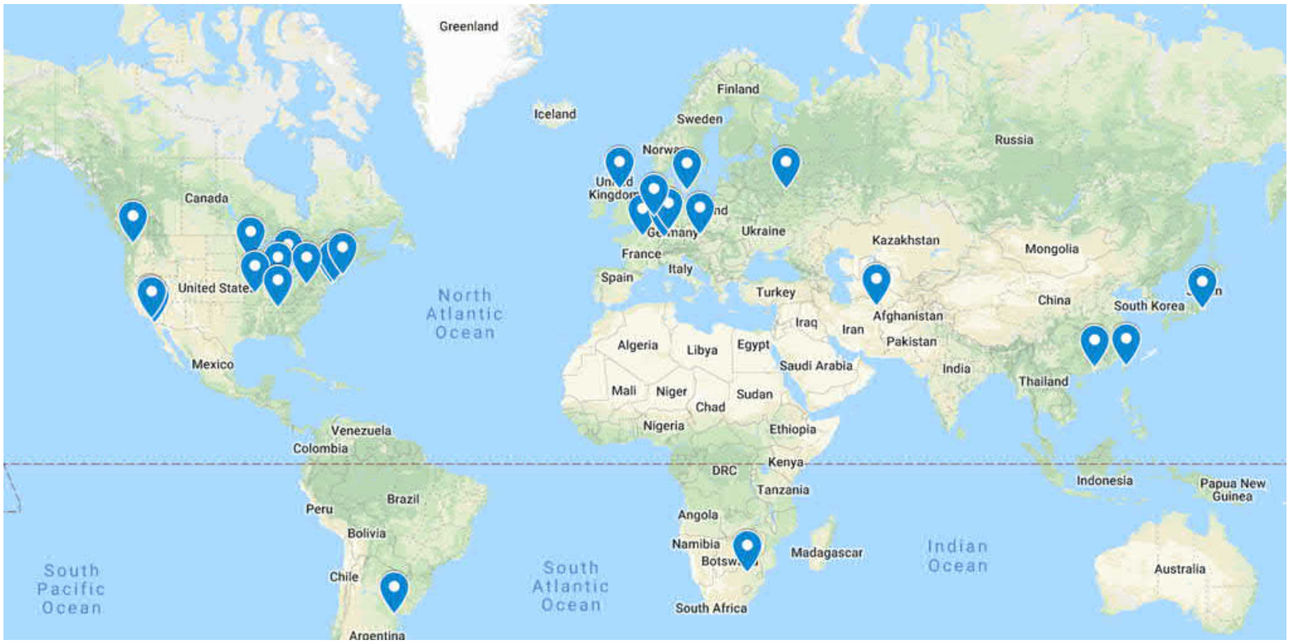


Figure 4. Geographic distribution of Grand Challenge 4 participants.

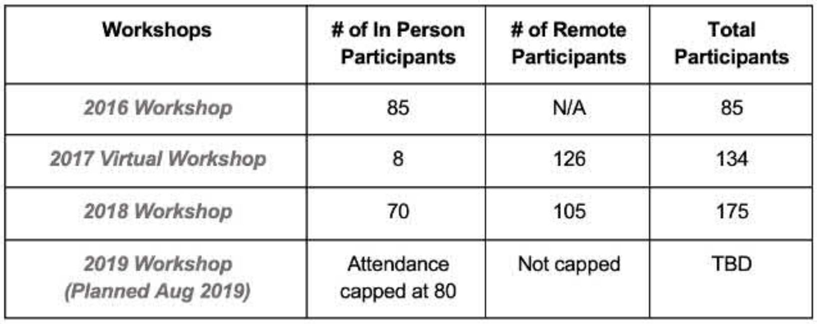
C.1.2 Continuous Challenges

A central purpose of this BTRR is the development, deployment, and use, of two continuous CADD challenges, one for pose-predictions (CELPP), and the other for affinity predictions or rankings (CELPP+). Unlike the Grand Challenges, which span several months each year, these high-volume, rolling challenge are not tied to a particular timeline. This is an advantage, because a new participant can join any time, but it risks losing the focus and excitement of a timed event. Therefore, we will put great effort into recruiting participants, as indeed already reflected by many Letters from prospective CELPP and CELPP+ participants in the Driving Biomedical Projects section. In addition, as noted above, we believe the Grand Challenges will continue to serve a highly visible and attractive on-ramp to the continuous challenges for newer participants. Accordingly, we will encourage Grand Challenge participants to take advantage of our workshops, trainings and hackathons (below), which will be designed to help them over the barrier of converting largely manual computational procedures to end-to-end, automated and containerized workflows suitable for the rolling challenges. Our partnership with MolSSI and BioSimSpace (again, see **Letters in Overall from Crawford and Michel**) will be of particular value in this regard.

C.2 Expand technology adoption through workshops, hackathons, and online training modules

C.2.1 Workshops

We will continue to hold the annual D3R workshop, either in person or virtually (e.g., a webinar), as these have proven to be a powerful means of engaging our community and have generated consistently strong participation (**Figure 1 & Figure 5)** . We have also used the workshops to extended the reach of our community engagement by partnering with the SAMPL initiative, which focuses on model systems for method development[17,18]. By intersecting these two groups, D3R enabled a highly productive interchange of ideas and progress among and between groups of researchers. In fact, post-workshop surveys (see below) made clear that there was great enthusiasm for this combined event, and that participants preferred the schedule to be organized so everyone could hear all the talks, rather than running D3R and SAMPL sessions in parallel. Therefore, we plan to continue holding joint D3R/SAMPL workshops (see **Letter of David Mobley in Overall**).



Workshops	# of In Person Participants	# of Remote Participants	Total Participants
2016 Workshop	85	N/A	85
2017 Virtual Workshop	8	126	134
2018 Workshop	70	105	175
2019 Workshop (Planned Aug 2019)	Attendance capped at 80	Not capped	TBD

Figure 5. In person and virtual D3R workshops. Remote participants joined electronically.

We piloted a virtual workshop in 2017, chiefly due to budget constraints. Well over a hundred participants called in from around the globe for this 5-hour webinar (**Figure 5**), and the post-workshop survey showed that this was a very popular option, particularly for our community members in far-away places. Therefore, when we hosted our next in-person workshop in 2018, we used the same infrastructure to simultaneously do a live broadcast of the event. This greatly extended the reach and impact of the meeting, as we had 70 in-person participants and 105 remote participants, bringing the total number of participants to its highest ever value (**Figure 5)**. This approach also allowed us to capture video of the talks, with speaker permission, and these are now linked on the D3R website (https://drugdesigndata.org//about/d3r-2018-workshop). We were very pleased at these outcomes and received much positive feedback from the community about these events. We therefore plan to continue supplementing our in-person meetings with virtual participation, and also holding some entirely virtual meetings. We plan to use GoToWebinar for these, because we have had good experiences with this technology, which, importantly, allows us to easily follow up with surveys for all attendees, and also gives us usage and attendance statistics, such as how long each user stayed online and if they were viewing actively. We will also continue to gather input from the community on how to improve such events, and indeed we close each meeting with a session that requests input from the community.

Workshops will typically comprise a combination of talks and open discussions on a range of topics, including participant talks about computational methods, results and conclusions; challenge overviews from D3R organizers and evaluators; perspectives from other community-members, such as industrial computational chemists; discussions of methods and standards for software development, data interchange, and workflow modularization; and discussions of data quality and evaluation metrics; and discussions of overall project directions. We will also consider joining add-on training sessions and hackathons (below) to the main workshops.

C.2.2 Software hackathons

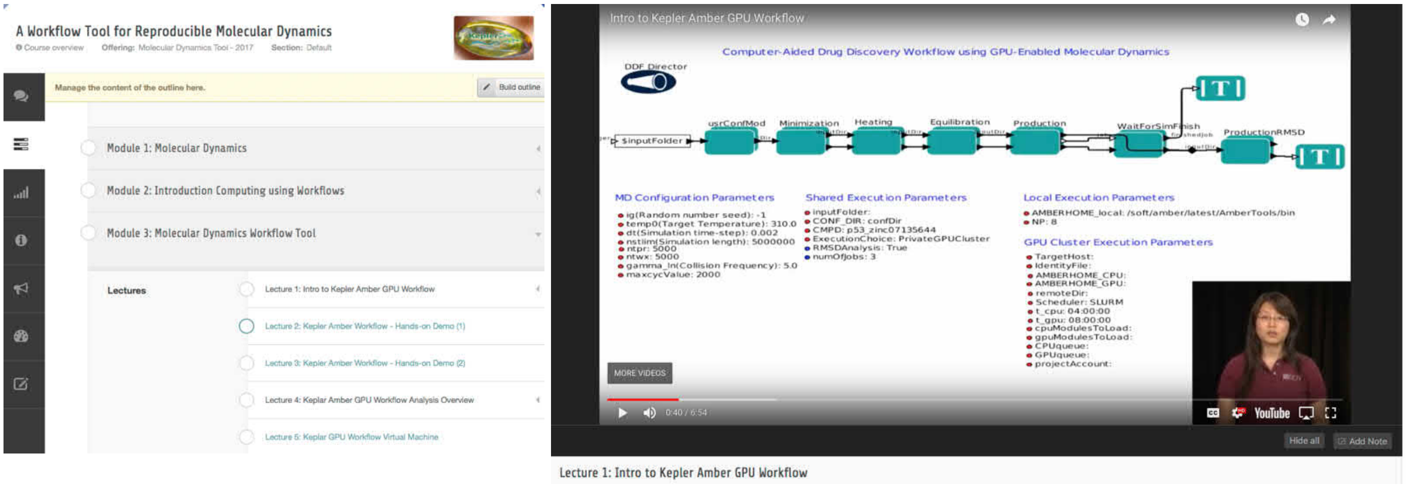
The impact of our proposed technology development depends on its adoption by the community. In particular, method developers will have to learn how to convert their largely manual docking and affinity calculations into end-to-end, containerized, workflows, that follow community-accepted standards for software implementation and data interchange. Accordingly, we plan to host, and have budgeted for at least intensive hackathons, held at least annually, to help participants take this critical step. These events will be developed along with two other major community projects.

One partner is the NSF-sponsored Molecular Sciences Software Institute (MolSSI, see **Letter of Support from Daniel Crawford, in Overall**), which aims to help molecular software scientists develop high-quality, sustainable software products[19,20]. Like D3R, MolSSI is concerned with ensuring that the community develops agreed-upon standards, in order to promote software and method interoperability and reuse. Therefore, they have agreed to partner with us specifically in the space of computer aided drug design methods and software development, co-organizing both hackathons and specialized workshop aimed at setting standards for data, methods, and evaluations in CADD. The other partner is the United Kingdom's flagship software development project of the Collaborative Computational Project in Biomolecular Simulations, BioSimSpace (see **Letter of Support from Julien Michel, in Overall**). BioSimSpace is developing workflows for binding free energy prediction, and has already formed a close link with D3R, including testing their workflows in our Grand Challenge 4[21–23]. They have agreed to be among our 'early adopters' group of DBP Investigators and also are keen to co-organize hackathons and workshops with us, both in the US and abroad. Both of these partnerships embody our strong collaborative network. We anticipate creating additional partnerships with other community allies as we progress into the new period, as opportunities arise.

C.2.2 Training Activities

We have received many requests from the research community to provide training in the CADD methods studied in this project, but the D3R's current funding mechanism does not include a training component. We are enthusiastic about including training activities in the present proposal. One general area which we believe will be well-received, and which flows naturally from the evaluative studies planned here, is training in best practices for pose-prediction and affinity ranking using CADD methods. The planned modularization of DBP workflows – for example, of docking into protein preparation, ligand preparation, conformational search, and rescoring – also lends itself naturally to the developing of corresponding training modules covering each of these steps. We are also eager to emphasize data science aspects of CADD. Doing so will enhance the capabilities of the existing and emerging CADD workforce and will help researchers take advantage of current technologies in their methods and applications. We will work with the community, including industry partners, to prioritize these and other potential training areas.

Trainings will be developed and led by D3R team-members and/or by trusted community experts interested in sharing their knowledge. Many CADD scientists in the industrial section are eager to share their knowledge with the next generation of trainees and they offer not only a translational view that will be new to some trainees, but also a valuable perspective of non-academic career options. These partners will also help us to disseminate our on-line training modules (below) through their professional networks and thus reach an even broader audience.



A Workflow Tool for Reproducible Molecular Dynamics

Course overview   Offering: Molecular Dynamics Tool - 2017   Section: Default

Manage the content of the outline here.

Module 1: Molecular Dynamics

Module 2: Introduction Computing using Workflows

Module 3: Molecular Dynamics Workflow Tool

Lectures

- Lecture 1: Intro to Kepler Amber GPU Workflow
- Lecture 2: Kepler Amber Workflow - Hands-on Demo (1)
- Lecture 3: Kepler Amber Workflow - Hands-on Demo (2)
- Lecture 4: Keplar Amber GPU Workflow Analysis Overview
- Lecture 5: Keplar GPU Workflow Virtual Machine

Intro to Kepler Amber GPU Workflow

Computer-Aided Drug Discovery Workflow using GPU-Enabled Molecular Dynamics

MD Configuration Parameters

- ig(Random number seed): -1
- temp0(Target Temperature): 310.0
- dt(Simulation time-step): 0.002
- nstlim(Simulation length): 5000000
- ntpr: 5000
- ntwx: 5000
- gamma_ln(Collision Frequency): 5.0
- maxcycValue: 2000

Shared Execution Parameters

- inputFolder:
- CONF_DIR: confDir
- CMPD: p53_zinc07135644
- ExecutionChoice: PrivateGPUCluster
- RMSDAnalysis: True
- numOfjobs: 3

Local Execution Parameters

- AMBERHOME_local: /soft/amber/latest/AmberTools/bin
- NP: 8

GPU Cluster Execution Parameters

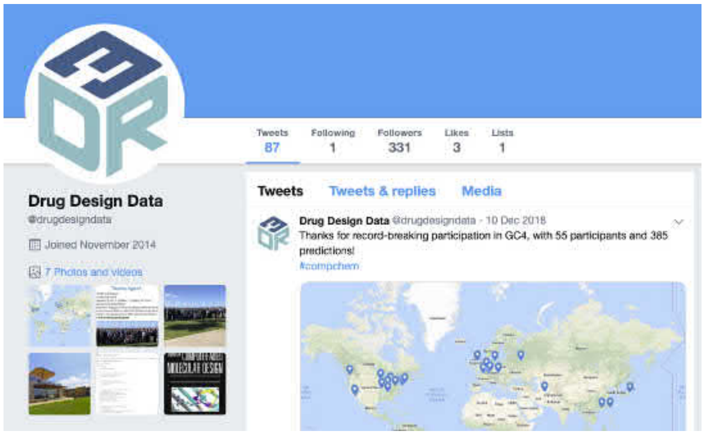
- TargetHost:
- IdentityFile:
- AMBERHOME_CPU:
- AMBERHOME_GPU:
- remoteDir:
- Scheduler: SLURM
- t_cpu: 04:00:00
- t_gpu: 08:00:00
- cpuModulesToLoad:
- gpuModulesToLoad:
- CPUqueue:
- GPUqueue:
- projectAccount:

Lecture 1: Intro to Kepler Amber GPU Workflow

Figure 6. Sample BBDTC training module demonstrating a Kepler workflow implementation of a molecular dynamics simulation with AMBER. Left: List of training modules on MD, force fields, computing with workflows, and the Kepler workflow tool. Right: a YouTube hosted lecture associated with the training module.

Initial training modules will be piloted at or in association with our annual workshop and then refined as needed. We will then capture these trainings and their associated materials, including workflows, with an innovative, collaborative training platform, the Biomedical Big Data Training Collaborative (BBDTC), which was co-developed by D3R investigators Altintas and Amaro under an NIH BD2K R25 Open Educational Resource award. The BBDTC is an online learning platform that includes big data science curricula, an open online course (OOC) framework, and a software toolbox for assembling hands-on tutorials (**Figure 6**). Containerized workflows can also run in this environment, so its concept aligns strongly with our proposed technology development work. Over the next five years, we will leverage and maintain this portal to extend the reach of our training activities. Although the funding period for the BBDTC grant has ended, the platform is fully featured and operates in production form.

C.3 Use electronic media to share information and interact with participants and the broader community of interest


Drug Design Data
@drugdesigndata
Joined November 2014
7 Photos and videos

Tweets 87 | Following 1 | Followers 331 | Likes 3 | Lists 1

Drug Design Data @drugdesigndata · 10 Dec 2018
Thanks for record-breaking participation in GC4, with 55 participants and 385 predictions!
#compchem

Figure 7. D3R will continue to use Twitter to disseminate timely information.

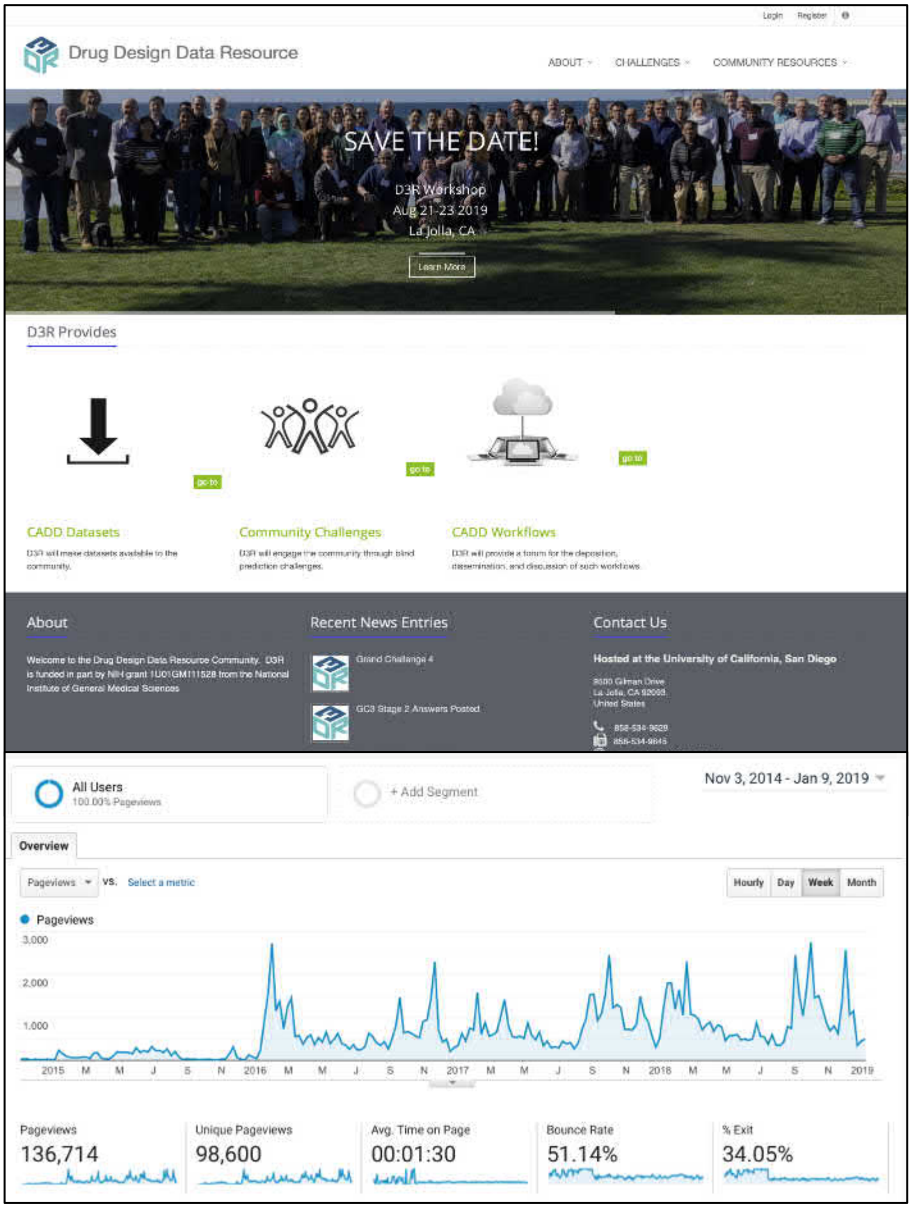
D3R will continue to make effective use of electronic media to disseminate information and expand our community of participants. D3R maintains email list of about 590 interested people, with sublists for individuals involved in specific events, such as challenges and workshops. People join the list by directly signing up for it on our website (below), or by signing up for a D3R event. The list is run with MailChimp, which allows us to avoid sending unwanted emails by placing an unsubscribe link at the bottom of each email message. For announcements to a broader segment of the computational community, we send occasional messages to the Computational Chemistry List (ccl.net). We will also continue to use the D3R Twitter account for this purpose, (**Figure 7**), as well as our personal Twitter accounts. Finally, the project has benefitted from the use of online survey tools, such as SurveyMonkey, to seek feedback about the content and organizational aspects of D3R workshops and the Grand Challenges.

We will furthermore maintain and expand D3R's comprehensive website (https://drugdesigndata.org), which is built on the SciCrunch platform (https://scicrunch.org). The website conforms to the P41 specifications provided in the PAR, and it has proven to be a highly effective communications vehicle, with steadily increasing user visits since our launch in 2015 (**Figure 8**). News and challenge updates are regularly updated on the landing page, and there is a dedicated area for Challenge-related activities, including the Grand Challenges and the preliminary CELPP implementation, as well as pages with datasets and challenge results, workshop information, D3R-related publications, links to the D3R GitHub page, and a page listing some of the lessons learned about CADD methods in the course of the project. We also partner with the SAMPL initiative[17,18] (see **Letter of David Mobley, SAMPL PI, in Overall**) and maintain descriptions of the SAMPL challenges as well.

C.3 Disseminate and publish data, technologies, and results

C.3.1 Data dissemination

Stable archiving and effective dissemination of data are central goals of this project, and indeed are a focus of technology development in **TRDP 3**, Data and Analytics. Several classes of data will be disseminated; these include the tens to hundreds of thousands of protein-ligand binding data to be curated in the course of this project, along with the existing ~1.5 million binding data currently served to the research public by BindingDB; the large numbers of computational predictions made by workflows in the CELPP and CELPP+ servers, along with basic statistical evaluations and compute-performance and computational provenance information; and the relatively modest volume of data and results associated with the Grand Challenges. The former data will all be archived in the D3RDB/BindingDB database discussed in **TRDP 3**. New cocrystal structures provided by industrial partners will be curated as needed and deposited in the Protein Data Bank by co-Investigator Stephen Burley's group, with cross-links to related data in D3RDB/BindingDB. As detailed in the TRDP proposals and also considered in Resource Sharing, D3RDB/BindingDB will be mirrored across at least two widely separated buildings, will be tagged with DOIs via the UC EZID facility, and will be shared openly via direct data downloads and the TRDP websites. The D3R team has extensive experience with data archiving and sharing through BindingDB, the PDB, and the D3R project itself.



D3R will make datasets available to the community.

D3R will engage the community through blind prediction challenges.

D3R will provide a forum for the deposition, dissemination, and discussion of such workflows.

Welcome to the Drug Design Data Resource Community. D3R is funded in part by NIH grant 1U01GM111528 from the National Institute of General Medical Sciences

Hosted at the University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093
United States
☎ 858-534-9629
856-534-9845

Figure 8. Home page of the current D3R website (top) and Google Analytics record of rising usage 2015-2019 (bottom).

C.3.2 Technology dissemination

Effective and transparent access to our technologies and data strengthens current collaborations, extends ties to new researchers who may subsequently become collaborators, and allows us to receive valuable input on technology and data functionality and potential enhancements. We frequently interact with the community by responding to requests for bug fixes, implementation of new assessment metrics, and general guidance on how the tools can be used most effectively.

One class of technology to be shared comprises the software methods and data standards we will collaboratively develop to establish a widely accepted and effective approach to constructing modular, containerized workflows. This is a set of guidelines and standards that developers may follow to create a piece of software that will integrate smoothly with the CELPP or CELPP+ frameworks. We anticipate that these standards will also be broadly useful as a basis for sharing CADD methods across the community and with other projects. We will work with MolSSI, BioSimSpace, and the community to capture these standards in documents, and to teach them at our workshops, hackathons and trainings (above).

Another class of technologies is the workflows themselves, several of which will have been developed by D3R, but most of which will have been developed by DBP Investigators (e.g., **DBPs 1-5, 7, 8, 10-13, and 15, 16**). Given required permissions, all workflows will be deposited in D3RDB/BindingDB in conjunction with their prediction results, the experimental data corresponding to the predictions, and workflow performance data, to create an integrated resource that allows researchers to obtain workflows fully annotated based on their accuracy in relation to documented experimental data, and their computational performance.

Finally, the software that runs CELPP and CELPP+ will be developed open-source on our GitHub repository, where it will be freely available for inspection and download. Software framework and community building activities will follow an iterative, requirements-driven cycle. Test cases will be established and applied as new versions are developed. We will hold weekly videoconferences to coordinate project activities, and will have periodic face-to-face team meetings to synchronize and expedite design and development. Code will be documented, and user guides will be written for the sake of continuity within the D3R project and to enable new users to adopt and use the code. We will adopt the practice of "continuous integration", which denotes merging and testing small changes to the code often, and will employ the Travis-CI build systems to automate testing after every update. Documentation also will be maintained on the GitHub site. Researchers interested in adopting this technology will be aided with deployment and use.

C.3.3 Dissemination of results

Peer-reviewed articles and conference talks are among the most effective means of scientific communication, and we will continue to use these traditional avenues to share the concepts and results of D3R research. First, we will continue work with Dr. Terry Stouch, Editor-in-Chief of the Journal of Computer-Aided Drug Design to release an annual special issue (**Figure 3**) about the D3R challenges (see **Letter of Terry Stouch in Overall**). As noted above, each issue has included an overview paper from D3R and typically ~20 articles by challenge participants. The overview papers alone are receiving about 30 citations per year. In addition, we anticipate that aspects of the CELPP and CELPP+ challenges will be more appropriate for other journals. Second, D3R investigators speak regularly at national and international meetings and conferences and will continue to use these opportunities to share information about D3R and recruit new DBP investigators. In addition, D3R will continue to sponsor participation by our graduate students, postdocs, and potentially staff at scientific meetings.

D Collaboration and Service

Although no formal collaboration service component is proposed during this initial phase, we will be open to consulting with researchers about related technologies and providing access to D3R resources as they may be available. Interested parties are free to contact D3R investigators and typically do so via our institutional email addresses or by using the project contact information on our website. The latter generates a message to multiple team members, who decide on and generate a response. As our technologies, and those of the DBPs mature, we will look for opportunities to establish Collaboration and Service relationships with researchers in the CADD community.

E. Technology Development Arc, the Expected Life of the Center and our Sustainability Plan

Based on our knowledge of the field and our experience over the past 4 years establishing D3R as an NIH U01 project, we expect that this project will be productive for 10-15 years as a NIH P41. Here we lay out our long-term vision for the Drug Design Data Resource.

During the first 5 years as a P41 BTRR, D3R will establish robust initial workflow-based technology frameworks for both CELPP and CELPP+. We also anticipate continuing our roughly annual Grand Challenges for at least the first few years, as a means of interaction with the community and recruiting new DBPs. We will help lead the national and international conversations around data standards and methods standards for CADD, along with other entities, including the NSF-sponsored MolSSI Institute (see **Letter of Daniel Crawford, Overall**) the UK's flagship simulation software project, BioSimSpace (see **Letter of Julien Michel, Overall**), and we will follow these standards in our operations. We will work with community members (e.g., **DBP Investigators**) to develop containerized, modular workflows for docking and affinity prediction and will develop the best performing open source workflows into accessible web services. We will recruit new DBP collaborators across all three Aims, adapting as needed in response to changes in the field.

In years 6-10, we anticipate operating in full production mode, accommodating many new workflows by migrating much of the compute load from our cluster to cloud resources. More work may need to be done breaking workflows into interoperable modules, but this will enable complex experiments that integrate methods from different groups into hybrid end-to-end workflows and lead to improved methods. These will support a growing community of data-science-embracing DBP users with drug-design applications. We anticipate that, over time, advanced computational methods developed in the DBPs using D3R technologies will be broadly and that their use, maintenance, and further development will be sustained by individual research grants and/or through commercialization. The CELPP and CELPP+ frameworks will also start becoming robust enough that some other groups may begin installing them on their own systems for in-house research. Note, however, that these will still depend on our data curation in TRDP 3. Thus, we will develop opportunities to control costs, such as reducing data curation costs and fostering sustainability by working with journals to require direct deposition of protein-ligand binding data into the public databases, and by collaborating with machine learning and natural language processing experts to ramp up reliable methods of partially or fully automating the extraction of benchmark data from articles and other documents—as indeed already begun in the current proposal. Finally, given community interest, it may be useful to take analogous approaches (including acquiring relevant data from industry) to address other outstanding challenges in CADD, such as predicting drug binding kinetics, membrane permeability, or toxicity.

Toward the end of years 6-10, we will assess the future role of D3R as then constituted. Much may have changed in the fields of CADD and software engineering, but we anticipate new opportunities to enhance CELPP and CELPP+ by integrating new software technologies, and to recruit DBP collaborators with innovative prediction methods. For example, assuming continued improvement in computer speed, we anticipate growing use of simulation-based free energy methods, possibly married to advanced machine-learning technologies. More exotic possibilities may also merit consideration, such as integrating our technologies with robotic chemical synthesis and testing, leading to integrated, iterative, self-learning structures that can evolve not only computer models but also targeted compounds with minimal human intervention.

References

1. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
2. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
3. Allen, W. J. *et al.* DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* **36**, 1132–1156 (2015).
4. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* (2019). doi:10.1007/s10822-018-0180-4
5. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* 1–20 (2017). doi:10.1007/s10822-017-0088-4
6. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **30**, 651–668 (2016).
7. Wagner, J. *et al.* Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *bioRxiv* (2018). doi:10.1101/469940
8. Yung-Chi, C. & Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099–3108 (1973).
9. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
10. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580–594 (2006).
11. Lavecchia, A. & Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **20**, 2839–2860 (2013).
12. Tembre, B. L. & Mc Cammon, J. A. Ligand-receptor interactions. *Comput. Chem.* **8**, 281–283 (1984).
13. Mortier, J. *et al.* The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes. *Drug Discov. Today* **20**, 686–702 (2015).
14. Gilson, M. K. & Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42 (2007).
15. Adcock, S. A. & McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **106**, 1589–1615 (2006).
16. Gumbart, J. C., Roux, B. & Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **9**, 794–802 (2013).
17. Bannan, C. C. *et al.* Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **30**, 927–944 (2016).
18. Yin, J. *et al.* Overview of the SAMPL5 host–guest challenge: Are we doing better? *J. Comput. Aided Mol. Des.* **31**, 1–19 (2017).
19. Dakka, J. *et al.* High-throughput binding affinity calculations at extreme scales. *BMC Bioinformatics* **19**, (2018).
20. Balasubramanian, V., Treikalis, A., Weidner, O. & Jha, S. Ensemble Toolkit: Scalable and Flexible Execution of Ensembles of Tasks. *ArXiv160200678 Cs* (2016).
21. Michel, J. & Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.* **24**, 639–658 (2010).
22. Granadino-Roldan, J. M. *et al.* Effect of set up protocols on the accuracy of alchemical free energy calculation over a set of ACK1 inhibitors. *bioRxiv* (2018). doi:10.1101/333120
23. Loeffler, H. H. *et al.* Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software. doi:10.26434/chemrxiv.6402425.v1

DRIVING BIOMEDICAL PROJECTS

SPECIFIC AIMS

The overarching biomedical goal of this project is to speed the development of new medications by helping researchers evaluate, advance, and utilize effective technologies for computer-aided drug design (CADD). The three TRD components are all designed to help developers of CADD software evaluate and improve their methods, and many DBPs will involve using these technologies for this purpose. Others will select operational, evaluated, CADD workflows from the D3RDB/BindingDB archive and put them to use in drug discovery projects.

The research community is strongly interested in methods of rigorously evaluating CADD technologies, as evidenced by high participation in the D3R Grand Challenges and D3R workshops. However, the Grand Challenges are quite limited in their ability to assess and compare computational methods, chiefly because of the modest number of test cases they use and their episodic – roughly annual – scheduling. The present project will generate test methods that are far more powerful, due to their continuous, high-throughput character. Our novel approach has already attracted commitments from many early-adopters wishing to participate as DBP collaborators, and, because the technologies to be developed are readily scalable, and the barrier to adding further DBPs will be low, we anticipate many more CADD-development DBPs to take shape, as the workflow technologies and standards mature and gain acceptance.

Furthermore, as new predictive workflows are developed and characterized, they will be made available, along with their detailed performance data, for review, download, and use, via the new D3RDB/BindingDB archive. These evaluated workflows thus will enable another class of DBPs, in which collaborators choose and apply selected workflows to their own drug discovery projects. Such validated workflows will also be available to DBPs for incorporation into even more comprehensive tool-chains. In a particularly innovative and far-reaching initial DBP in this group, D3R-evaluated workflows will be used to provide feedback on novel candidate ligands designed by citizen-scientists in a drug-design “game” analogous to the protein design game, FoldIt.

In summary, we aim to build and support collaborative DBPs with the following Aims:

Aim 1. Take advantage of the CELPP TRD 1 to advance pose-prediction technologies. In addition to providing the continuous evaluation technology, D3R contribution to these collaborations will include helping with data exchange and technology standards, assistance with workflow containerization, maintenance of operational workflows, and provision of raw data and analytics.

Aim 2. Take advantage of the CELPP+ TRD 2 to advance affinity calculation/ranking technologies. D3R will deploy, execute, evaluate, and disseminate the results of continuous blinded challenges for binding affinity predictions. As part of this effort, we will and-in-hand with our DBP investigators to help encode their workflows into containers that can be continually interrogated with our data streams developed in **TRDP3**.

Aim 3. Put validated workflows identified through D3R activities to use in drug design projects. D3R will provide a unique service of annotating workflows with information about their accuracy and speed across a wide range of challenge cases. Workflows that perform well by relevant metrics may be selected by drug design teams for use in their projects. This Aim also encompasses an innovative, extensible, crowd-sourced approach to drug design.

Across all three Aims, we will also collect feedback regarding all three TRDPs from DBP collaborators and direct it to the TRD teams to improve and refine their technologies and user-interfaces. Collaborating with the DBP investigators will test the new technologies developed here and will drive corrections and advances.

RESEARCH STRATEGY

A Significance

D3R aims to create new technologies and provide new and improved access to data in order to dramatically advance methods in computer aided drug design (CADD). Our focus in this first proposed project period is on two aspects of CADD: pose prediction (docking), and ligand ranking (scoring or binding affinity prediction). We choose to focus on these two methodological aspects of CADD because they are essential to most drug discovery and design programs, our earlier efforts give us a strong footing on which to evolve the technological development, and we have established a large cadre of participants in the community, as indicated by our extensive network of driving biomedical projects (DBPs) and collaborators. The DBPs prioritize our technology development and motivate our creation of new methods and tools.

D3R is the only resource of its kind in the world, and over the past 5 years, under an NIH U01 award since 2014, **D3R has established itself as the global hub for the CADD methods development community**. Our blinded community challenges, a longstanding component of D3R and, to some extent, its predecessor CSAR, have episodically brought together the CADD community to challenge methods with blinded datasets. Through these activities, D3R has shown a strong ability to bring academic method developers and industrial and academic CADD methods / software users and advance the field. D3R's success in growing the global research community over the past 5 years highlights the centrality of these methods to biomedical research worldwide. Because we serve this global community through our efforts, our DBP investigators come to us from the United States, as well as many other countries around the globe.

B Overarching Aims of DPB Collaborations

A common need for the CADD community, and thus of all our DBPs, is the need for a new technological framework to advance pose prediction and ligand scoring methods. The concept of developing this much-needed technology is at the heart of D3R. CADD methods touch many different research programs in different ways, so the activities of D3R impact both fundamental and applied components of the biomedical research enterprise.

To provide a high-level perspective and succinctly describe how our activities are driven by various stakeholder needs, we have organized our initial DBP portfolio into three broad Aims:

- **Aim 1: Enabling Evaluation and Advancement of Pose Prediction**
- **Aim 2: Enabling Evaluation and Advancement of Ligand Affinity Prediction or Ranking**
- **Aim 3: Applications of Evaluated CADD Methods & Innovative Uses of D3R Technologies**

Note that a number of projects will actually advance multiple aims. For example, one research group may be developing software that combines pose prediction and affinity ranking.

It is also worth emphasizing that Aim 1 and Aim 2 investigators nearly all have their own ongoing collaborations to use their software in various biomedically relevant disease areas. In addition, by providing executable, containerized workflows, which will be archived and thoroughly documented in the Data and Analytics TRDP, each DBP collaborator makes their tools and codes available for broad use across the drug-design community. Thus, D3R is able to reach a huge audience of biomedical researchers; for example, just one of the docking codes whose methods development we enable, AutoDock Vina, has over 10,000 registered users! Furthermore, the experimental data collected in the Data and Analytics **TRDP 3** will be used worldwide by medicinal chemists, computational chemists, machine learning experts, systems biologists, and even clinicians. Aim 3 investigators, who may be broadly characterize as "end users", exemplify how the technological frameworks we enable at D3R impact active drug discovery research programs, both in academic and industrial (pharmaceutical) settings. This includes the development of optimized end-to-end workflows and rigorous assessment of methods against real world drug targets, as well as highly innovative extensions to our current and proposed technology products for use in other settings.

B.1 Aim 1: Enabling Evaluation and Advancement of Pose Prediction Methods

B.1.1 Background and Significance: Reliable computational tools to predict the poses of small molecules in the binding site of targeted proteins would allow drug designers to quickly identify the chemical sites on a ligand that are appropriately positioned and angled to direct a new substituent into a concave subsite of the binding pocket of the targeted protein, and thus potentially to design compounds of higher affinity. In addition, accurate pose prediction, or docking, is a key first step in structure-based methods to predict ligand-protein affinities (Aim 2). It can also provide insight into biomolecular mechanisms; for example, knowing precisely how an enzyme binds its substrate is essential to understanding its catalytic mechanism and biological activity. As a consequence, researchers worldwide are working to solve the docking problem. However, progress toward this goal is impeded by the fact that it is difficult to determine unambiguously whether a methodological change represents an advance, or even to compare two distinct methods on a solid footing and thus to select a method for a given application. One reason is that different methods are typically tested against different sets of protein-ligand complexes, so there is no consistent basis for comparison. More important, however, is that reported tests of docking methods typically are based on publicly available protein-ligand cocrystal structures, and that many of the same publicly available test cases are used over and over. Such studies are suboptimal, because they risk unintentional bias and because structures in the test set might have been used previously in training the docking algorithms.

Widespread recognition of the need for an improved paradigm for evaluating docking methods motivated broad participation in the D3R Grand Challenges and in our preliminary version of the CELPP continuous, blinded pose-prediction challenge. The present project will take objective, large-scale, evaluation of pose-prediction algorithms to a new level, enabling rigorous comparisons among methods down to the component level, and allowing developers to engage in an efficient, iterative cycle of evaluation, development, and re-evaluation. We anticipate that some methods will even automate this learning process and become more accurate over time without human intervention. The large-scale impact will be to make increasingly powerful pose-prediction methods available to the biomedical research community. Furthermore, by moving to containerized workflows and archiving these, along with their results for thousands of defined protein-ligand systems, we will create a novel resource of methods annotated by their performance. Thus, developers will be able to reproduce results and drug designers will be able to make informed technology choices.

The Aim 1 DPBs are projects that will take advantage of **TRDP 1**, primarily, to evaluate and advance the technology of protein-ligand pose predictions.

B.1.2 Approach: D3R's approach to overcoming these challenges is embodied in the CELPP challenge[1], which is described in **TRDP 1** and supported by **TRDP 3**. CELPP uses the Protein Data Bank's (PDB) weekly publication of a list of structures slated for imminent release into the public domain as the basis for a weekly pose-prediction challenge. In the current, preliminary implementation of CELPP, developers receive a weekly data package from us, then, over the course of ~ 5 days, generate pose predictions on their own machines and transfer a data package back to us that contains the ligands docked into the receptor structures. We evaluate the prediction and make the evaluation results available via the web. **By using a previously unexploited flow of PDB data, D3R has already provided a transformational advance for docking methods developers, who now have facile, consistent access to a large source of relevant structural data. Developers can now receive new feedback on their methods on a weekly basis and use this flow of evaluative data as a basis for continual improvement.** thus, in just the first 70 weeks of CELPP, D3R has provided the field over 3,000 blinded ligand-receptor docking opportunities. This number exceeds, by an order of magnitude, all the docking cases brought to the community over the past 9 years using the existing paradigm of episodic challenges. We have worked with several early adopters to develop this initial technology: David Koes (University of Pittsburgh, **DBP 8**), Xiaoqin Zou[2] (University of Missouri, **DBP 16**), and John Bohmann (Southwest Research Institute, **DBP 1**). These participants have provided frequent, rapid, and valuable feedback as we developed our technology platform. This class of early adopters are analogous to DBP investigators under the P41 mechanism.

Although the preliminary implementation of CELPP has already been transformational, much work is still needed to develop analytic tools that will provide insight into the results. In addition, the advent of new technologies for defining workflows and packaging them in easily shared containers (e.g., Docker and Singularity) have created important to opportunities to support reproducibility and dissemination of methods. As detailed in **TRDPs 1 and 3**, D3R seeks to develop a server on which Aim 1 DBP investigators will deposit their containerized workflows. Then, instead of D3R sending out packages of data to be evaluated for pose prediction success, D3R will maintain this server of codified complex workflows for docking, interrogating them in an automated fashion on a weekly basis. DBP investigators who partner with us have much to gain from their involvement and efforts to transform their code to automated workflows. For example, they will have access to our hardware infrastructure for the execution of their methods; to fair, reliable, and trustworthy evaluations of their methods; and a comprehensive archive of data in **TRDP 3** which will allow them to compare with and benefit from learning about other methods.

The developers of many of the world's leading docking methods have committed to participate in the development and use of this technology; examples include HADDOCK[3,4] (**DBP 2**), AutoDock Vina[5–7] (**DBP 5**), MDock[2,8,9] (**DBP 16**), Rhodium[10] (**DBP 1**), MolSoft[11,12] (**DBP 13**), GNINA and SMINA[13,14] and AnchorQuery[15] (**DBP 3**), MathParm[16] (**DBP 15**), and DOCK[17] (**DBP 12**). We will collaborate with these outstanding teams to apply the containerized workflows concept to each of their methods. This will not only dramatically improve their ability to evaluate the performance of their methods but will also support the training goals of this project while making many software packages easier to distribute and use.

Our approach also includes planned hackathons designed to help developers implement their codes as containerized workflows to be mounted on the D3R server. In some cases, we will work with other organizations to host these. For example, the NSF-sponsored MolSSI project, which aims to standardize and make molecular simulation software more robust, has agreed to partner with us in this space and in the cosponsorship of hackathons and workshops (see **Letter of Support in Overall from Daniel Crawford**). In addition, the United Kingdom's flagship project for biomolecular simulation software development, BioSimSpace, has also agreed to partner with us for workshops and hackathons (see **Letter of Support in Overall from Julien Michel**). Such efforts will be critical to the effective development, deployment, and sustainability of the developed technologies.

B.1.3 Driving Relationships of Aim 1 DBPs to TRD Projects: The DBP investigators are important drivers of our developing technologies because they inform us about particular requirements they may have that we need to accommodate; they help us develop community standards for docking codes, which will assist in the planned development of interoperable modules; and they test our new technologies and give us rapid and honest feedback about our technology performance and capabilities. Our DBP investigators in turn provide suggestions for technological improvement, and ultimately become representative consumers for what we help develop.

B.1.4 Impact of Enabling Evaluation and Methods Development for Pose Prediction: The impact of helping the developers of docking codes to improve their software is extensive, because each of the investigators that we work with has a community of users, each pursuing their own grant-funded biomedical projects. As a consequence, thousands of labs are expected to benefit. For a more detailed description of some of these efforts, please see DBP Aim 3.

B.2 Aim 2: Enabling Evaluation and Advancement of Methods for Ligand Affinity Prediction or Ranking

B.2.1 Background and Significance: In many drug discovery projects, the ability to reliably estimate the affinity of a candidate ligand for the targeted protein, or at least to assess whether it will bind with higher affinity than previously characterized compounds, would be of enormous value. This challenge can arise in multiple stages of the project. Initially, one would like to computationally scan a chemical library for hit compounds -- relatively weak binders, which serve as initial toeholds in a climb to higher affinity. This hit-to-lead process also would be accelerated by effective computational prediction methods, as would the subsequent stage of lead-optimization, in which a lead compound is further modified to improve affinity while also meeting orthogonal criteria related for example, to bioavailability, half-life, and distribution. It is worth noting that, although accurate pose-prediction will help with structure-based ligand design (Aim 1), going from poses to affinities is a major additional step, and an unsolved problem. Moreover, some affinity prediction methods are not structure-based, but instead use machine-learning methods trained on the identities of ligands and their targets, and the associated experimental affinities. Such methods are independent of pose-prediction technologies, and instead rely on large archives of affinity data, such as BindingDB in **TRDP 3**. (There are also methods that combine structural modeling and machine learning, and thus require pose prediction and historical binding data.) However, the affinity prediction challenge has in common with docking the need for a tool to objectively and statistically meaningfully evaluate and compare methods on an ongoing basis. Indeed, as in the case of docking, rigorous evaluation of affinity prediction methods has long been extraordinarily difficult due to the lack of blinded test cases, the testing of different methods on different cases, and the modest size of typical test sets. In particular, although the D3R Grand Challenges which have the merit of enabling many groups to test their methods on a shared set of blinded data, the modest size of the test sets means the results are more anecdotal then statistically significant.

Furthermore, predicting binding affinity can be a significantly more complicated task than predicting a ligand binding pose. Indeed, structure-based affinity methods have in most cases quite complicated end-to-end workflows, often with 50 or more key decision points left as options to the user. Thus, the current paradigm of submitting and evaluating only a single set of predictions (e.g., in terms of the Kendall's tau statistic for ligand ranking) severely limits the interpretability of the results, if only because another user cannot be sure of running the method in exactly the same way. It is thus essential to encode each method into a workflow where the key decision points are selected and controlled for while the method is evaluated over many test sets. It also then becomes possible to systematically determine the sensitivity of the results to the various decisions and parameters. Thus, the encoding of prediction methods into complete end-to-end workflows, as planned for **TRDP 2**, will allow each method to be rigorously evaluated, will help define optimal usage patterns, and will afford deeper insights into its strengths and limitations.

The Aim 2 DPBs will be projects that take advantage of **TRDPs 2 and 3** to evaluate and advance the technology of protein-ligand affinity predictions and thereby speed and facilitate the discovery of new medications.

B.2.2 Approach: D3R is proposing a novel and much needed approach to generating continuous, effectively blinded, challenges, in the space of protein-ligand affinity ranking and prediction. This approach relies on the development of a dedicated server that will host containerized workflows, and which will run on a continuous basis as data that D3R curates from articles and patents or either brokers from industrial partners are fed into the server. The results from the workflows and the workflows themselves then will be archived, along with the data themselves, in our D3R/BindingDB database and shared through our web portal. We, and the majority of the community, are convinced that while the development of containerized workflows for binding prediction methods will require a substantial amount of effort on behalf of the community, it is an effort worth undertaking. Thus, **working with developers and users in the CADD community, we seek to transform the evaluation and advancement of binding affinity methods as we have already begun to do for docking.** Many methods developers in this space have already confirmed their support of this initiative by promising to commit effort and energy into working alongside us as early adopters for the development of this new technology. These include the developers of AutoDock and AutoDock VINA[5–7] (**DBP 5**), GNINA and SMINA[13,14] (**DBPs 3 and 8**), MDock[2,8,9] (**DBP 16**), MolSoft[11,12] (**DBP 13**), DOCK[17] (**DBP 12**) as well as free energy method developers, Zoe Cournia[18,19] (**DBP 3**), Julien Michel[20–23] (**DBP 10**), Bogdan Iorga[24–29] (**DBP 7**), and David Minh[30] (**DBP 11**). Note that the hackathons, training events, and standardization efforts to be coordinated with the NSF-sponsored MolSSI project and the UK's BioSimSpace, (see Aim 1 Approach) will also be fully applicable to the Aim 2 DBPs. Through these coordinated efforts, D3R stands to dramatically advance capabilities of ligand ranking and binding affinity prediction methods.

B.2.3 Driving Relationships of DBPs to TRD Projects: All DBPs in this thematic area will contribute to technology development across **TRDPs 2 and 3**; DBPs where pose prediction is part of the affinity prediction may also contribute to **TRDP 1**. D3R will use input from DBP teams when designing our workflow standards and framework, and our incorporation of their containerized workflows into our framework will drive corrections and improvements to our technology. As we flow data through the workflows and evaluate, archive, and disseminate the results, feedback from the DBP teams will guide our procedures, data standards, file formats, evaluation approaches, and archiving and dissemination practices. For example, the design of the D3RDB/BindingDB archive in **TRDP 3** will be driven by analysis of the various DBP workflows and the data they use and generate, and by the requirements of data-driven workflows for access to data in the local BindingDB archive in the workflow framework; while **TRDP 2** will be driven, for example, by analysis of the potential modularity of the workflows we receive and by the range of their computational needs. In addition, inconveniences and failure modes in all TRD Projects will be identified and addressed as they may arise. Thus, the DBPs will drive both design and ongoing improvement and hardening of our technology platform, so they become more accessible to new DBP collaborators and more valuable to the research community as a whole.

B.2.4 Impact of Evaluation and Advancement of Methods for Ligand Affinity Prediction or Ranking

The impact of advancing methods for protein-ligand affinity predictions will directly impact tens of thousands of scientists in the community of users working on drug discovery projects in various academic and industrial settings worldwide; some of them are discussed below in DBP Aim 3.

B.3 Aim 3: Using D3R-Evaluated Methods for Drug Discovery

B.3.1 Background and Significance: In the course of this project, containerized workflows will be added to the CELPP (**TRD 1**) and CELPP+ (**TRD 2**) frameworks, test data from various types of protein-ligand systems will be flowed through them, and the workflow settings and results will be used to populate the workflow component of the D3RDB/BindingDB archive (**TRD 3**). Thus, **another valuable product of this project will be a collection of containerized workflows, annotated with performance data on various types of systems, and ready for use on active drug discovery projects.** (Note that, although we will encourage developers to limit their workflows to open-source code, some workflows may incorporate a module that requires purchase of a third-party license. In such cases, other researchers wishing to use the workflow will still be able to do so if they obtain the required license.) In DBP Aim 3, we will partner with outside researchers seeking to use these evaluated methods, and these collaborations will drive our successful development of a new paradigm of shared, evaluated workflows for drug discovery. We foresee at least two different ways in which D3R workflows will be applied to drug discovery projects.

First, and most straightforwardly, we will work with individual labs. Researchers in Alzheimer's Disease[31,32] (Dr. Steven Wagner, UC San Diego, **DBP 14**), as well as cancer and HIV[33] (Dr. Dan Harki, U. Minnesota, **DBP 6**) will be early collaborators on identifying useful D3R workflows and adapting them for use in NIH-funded drug design projects. These initial projects will enable us to identify and implement effective practices for integrating our evaluated workflows into the applied setting, as considered in the Approach section below (Individual Research Groups).

Second, we will collaborate with software developers on the selection of D3R workflows and their integration into larger tool-chains designed for use by many labs for a variety of projects. This class of Aim 3 DBP is exemplified by **DBP 9**, a highly innovative collaboration with Prof. Jens Meiler (Vanderbilt U) and Dr. Andreas Bergner (Boehringer-Ingelheim (BI), Vienna, Austria) that will recruit citizen-scientists from the public to contribute to the design of ligands in real-world drug discovery projects. This project, preliminary named 'Drug Discovery @ Home' may be considered the ligand design cousin of the well-known FoldIt protein design game. Here, the role of D3R will be to set up one or more workflows to provide "Level 2" evaluations of candidate ligands designed by members of the public using a faster, simpler "Level 1" scoring function built in to the interactive program running on the user's computer.

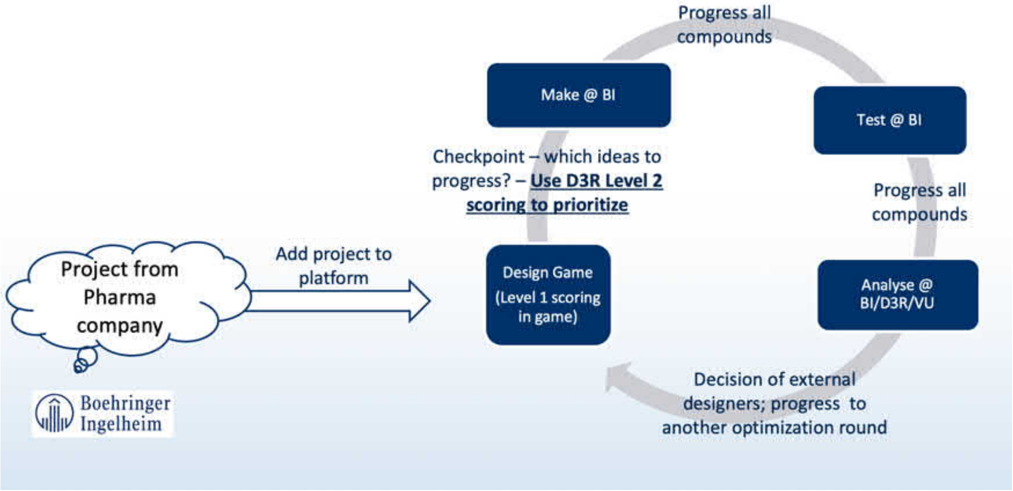

Progress all compounds
Make @ BI
Test @ BI
Checkpoint – which ideas to progress? – Use D3R Level 2 scoring to prioritize
Progress all compounds
Project from Pharma company
Add project to platform
Design Game (Level 1 scoring in game)
Analyse @ BI/D3R/VU
Decision of external designers; progress to another optimization round
Boehringer Ingelheim

Figure 1: DBP with Meiler (Vanderbilt) and Bergner (Boehringer Ingelheim, BI) embodies a novel use case of our developing technology. A 'game' to optimize a compound in the binding site of a drug target is launched from BI into the game; citizen scientists build and design the molecule, with level-1 scores being generated inside the game console; some top hundreds of compounds from level-1 scoring move on to second level scoring, which will utilize top performing workflows at D3R. The top compounds are then synthesized and tested at BI; analysis occurs across the three sites and then, if desired, additional rounds of optimization will ensue.

B.3.2 Aim 3 Approach: For individual research groups, such as those of Wagner (**DBP 14**) and Harki (**DBP 6**), we will work with the DBP labs to help them understand the information in the D3RDB/BindingDB data archive and use it to identify the most suitable workflow for their project. We will then help them integrate the selected workflow with their preferred drug-design practices. For example, the choice of workflow may depend on the family of the protein target and the number of compounds to be studied; and, depending on the computational requirements of the method, the workflow may be used as a remote server hosted by D3R, or we may help the lab to install and use the workflow on their own system. In some cases, it may be practical to help a local computational chemist integrate a D3R workflow with a molecular modeling interface already in use in the lab.

For collaborations with software developers, as in the Vanderbilt/BI collaboration, we will, similarly, assist in the selection of a suitable workflow and assist with integrating it with the larger tool-chain. For example, we will explain the data standards and interchange formats. Again, it may be helpful for us to host a workflow on the D3R server for access by a webservices API. This may be particularly useful for the crowd-sourced drug design application, where the typical citizen scientist cannot be tasked with installing and running an additional application. In this setting, we will need to work with the developers to ensure either that response times are reliably fast and/or to support a user interface that accepts delayed computational results in a graceful manner.

B.3.3 Driving Relationships of Aim 3 DBPs to TRD Projects: Working with these DBPs will help us ensure that the needs of the drug discovery community are met by the workflows we and our Aim 1 and Aim 2 DBP collaborators. We will also learn effective practices for making these software products work for applied users. For example, if a workflow has a very large number of adjustable parameters but works best with only certain settings, we will develop ways of preventing users from modifying these settings, unless there is a specific reason to do so, while leaving other, less critical, settings exposed to the users. We will also learn more about when/whether it is best for users to run the workflow on their own computers, on the D3R server, or, potentially, on a commercial cloud. As we become more experienced with effective practices and as more evaluated workflows come on line, we will be able to provide increasingly efficient and effective assistance to new DBP collaborators doing drug design.

B.3.4 Impact of Using D3R-Evaluated Methods for Drug Discovery: This project will develop a completely novel resource, an archive of containerized, downloadable, executable, CADD workflows, each annotated with information regarding its accuracy and computational performance when applied to an unprecedently large blinded test set (**TRDP 3**). As developers use CELPP (**TRDP 1**) and CELPP+ (**TRDP 2**) to improve their methods, the accuracy of these workflows will improve with time. The Aim 3 DPBs will take advantage of this novel resource in drug discovery projects and advanced applications, while providing feedback to D3R that will enable us to maximize the impact of the archive and its workflows. This aspect of D3R has enormous potential to accelerate real-world drug discovery efforts across all biomedical research areas.

C. DBP Recruitment, Selection, Monitoring and Turnover. As evidenced by the table of DBPs, D3R works diligently to cultivate high impact collaborative projects. We accomplish this by networking with top investigators all over the US and world, and through strategic outreach when necessary. DBPs are recruited in three ways: (1) we harvest / develop new collaborations at our annual workshops, (2) potentially new DBP collaborators may contact us via the D3R website or direct email to any D3R personnel, and (3) we reach out directly to specific highly engaged community members to develop a partnership. Enabling the DBPs is central to this project, so they will be constantly recruited, monitored, and developed. As the DBPs and the technologies they inspire us to develop mature, the DBPs may be moved to Collaboration and Service Projects (CSPs), may be retired (potentially to continue elsewhere as a non-collaborating project), or may stimulate development of entirely new projects. Importantly, our technologies will highly scalable, so we anticipate being very open to almost any participant, given adequate computer resources. Note that technologies will have capacity to go to commercial cloud providers if/when that proves cost-effective.

D. Plans to Strengthen DBP Collaborative Relationships. D3R has established itself as a highly community-centered resource, based on the participatory nature of the Grand Challenges, regular workshops – both face-to-face and web-based – and ongoing community engagement by judicious use of additional avenues such as email and social media. We plan to continue these mechanisms, making a point of involving current and potential DBP investigators and also inviting these researchers to UC San Diego for focused discussions about shared projects. Recruitment of particularly active and interested D3R participants investigators has been a cornerstone of our success to date in the development and implementation of our new technologies, such as CELPP. We will also work with other organizations to co-organize additional workshops and hackathons that will involve DBP investigators. Venues will include the large international meetings, such as the American Chemical Society national meetings and Gordon Research Conferences, as well as smaller venues, such as planned workshops with MolSSI and BioSimSpace, in order to bring multiple experts together to discuss technology development requirements for a particular CADD technology and/or the use of developed technology in a broader context. In D3R and in other community-oriented projects, we have found these mechanisms to be useful and efficient ways of communicating with others, enhancing productivity, and creating and strengthening collaborations.

References

1. Wagner, J. *et al.* Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *bioRxiv* (2018).
2. Xu, X., Ma, Z., Duan, R. & Zou, X. Predicting protein–ligand binding modes for CELPP and GC3: workflows and insight. *J Comput Aided Mol Des* (2019).
3. Kurkcuoglu, Z. *et al.* Performance of HADDOCK and a simple contact-based protein–ligand binding affinity predictor in the D3R Grand Challenge 2. *Journal of Computer-Aided Molecular Design* **32**, 175–185 (2018).
4. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *Journal of Computer-Aided Molecular Design* (2019).
5. Perryman, A. L., Santiago, D. N., Forli, S., Santos-Martins, D. & Olson, A. J. Virtual screening with AutoDock Vina and the common pharmacophore engine of a low diversity library of fragments and hits against the three allosteric sites of HIV integrase: participation in the SAMPL4 protein–ligand binding challenge. *Journal of Computer-Aided Molecular Design* **28**, 429–441 (2014).
6. Mobley, D. L. *et al.* Blind prediction of HIV integrase binding from the SAMPL4 challenge. *Journal of Computer-Aided Molecular Design* **28**, 327–345 (2014).
7. Gallicchio, E. *et al.* Virtual screening of integrase inhibitors by large scale binding free energy calculations: the SAMPL4 challenge. *Journal of Computer-Aided Molecular Design* **28**, 475–490 (2014).
8. Duan, R., Xu, X. & Zou, X. Lessons learned from participating in D3R 2016 Grand Challenge 2: compounds targeting the farnesoid X receptor. *J Comput Aided Mol Des* **32**, 103–111 (2018).
9. Xu, X., Yan, C. & Zou, X. Improving binding mode and binding affinity predictions of docking by ligand-based search of protein conformations: evaluation in D3R grand challenge 2015. *J Comput Aided Mol Des* **31**, 689–699 (2017).
10. Jaundoo, R. *et al.* Using a Consensus Docking Approach to Predict Adverse Drug Reactions in Combination Drug Therapies for Gulf War Illness. *International Journal of Molecular Sciences* **19**, 3355 (2018).
11. Lam, P. C.-H., Abagyan, R. & Totrov, M. Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach. *Journal of Computer-Aided Molecular Design* **32**, 187–198 (2018).
12. Lam, P. C.-H., Abagyan, R. & Totrov, M. Hybrid receptor structure/ligand-based docking and activity prediction in ICM: development and evaluation in D3R Grand Challenge 3. *Journal of Computer-Aided Molecular Design* (2018).
13. Sunseri, J., Ragoza, M., Collins, J. & Koes, D. R. A D3R prospective evaluation of machine learning for protein-ligand scoring. *J Comput Aided Mol Des* **30**, 761–771 (2016).
14. Sunseri, J., King, J. E., Francoeur, P. G. & Koes, D. R. Convolutional neural network scoring and minimization in the D3R 2017 community challenge. *J. Comput. Aided Mol. Des.* (2018).
15. Koes, D. R., Dömling, A. & Camacho, C. J. AnchorQuery: Rapid online virtual screening for small-molecule protein–protein interaction inhibitors. *Protein Science* **27**, 229–232 (2018).
16. Nguyen, D. D. *et al.* Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *arXiv:1804.10647 [q-bio]* (2018).
17. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA* (2009). doi:10.1261/rna.1563609
18. Athanasiou, C., Vasilakaki, S., Dellis, D. & Cournia, Z. Using physics-based pose predictions and free energy perturbation calculations to predict binding poses and relative binding affinities for FXR ligands in the D3R Grand Challenge 2. *Journal of Computer-Aided Molecular Design* **32**, 21–44 (2018).
19. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling* **57**, 2911–2937 (2017).
20. Papadourakis, M., Bosisio, S. & Michel, J. Blinded predictions of standard binding free energies: lessons learned from the SAMPL6 challenge. *Journal of Computer-Aided Molecular Design* **32**, 1047–1058 (2018).
21. Mey, A. S. J. S., Jiménez, J. J. & Michel, J. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *Journal of Computer-Aided Molecular Design* **32**, 199–210 (2018).

22. Bosisio, S., Mey, A. S. J. S. & Michel, J. Blinded predictions of distribution coefficients in the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **30**, 1101–1114 (2016).
23. Bosisio, S., Mey, A. S. J. S. & Michel, J. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **31**, 61–70 (2017).
24. Chaput, L., Selwa, E., Elisée, E. & Iorga, B. I. Blinded evaluation of cathepsin S inhibitors from the D3RGC3 dataset using molecular docking and free energy calculations. *Journal of Computer-Aided Molecular Design* (2018).
25. Selwa, E., Kenney, I. M., Beckstein, O. & Iorga, B. I. SAMPL6: calculation of macroscopic pKa values from ab initio quantum mechanical free energies. *Journal of Computer-Aided Molecular Design* **32**, 1203–1216 (2018).
26. Kenney, I. M., Beckstein, O. & Iorga, B. I. Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field. *Journal of Computer-Aided Molecular Design* **30**, 1045–1058 (2016).
27. Selwa, E., Martiny, V. Y. & Iorga, B. I. Molecular docking performance evaluated on the D3R Grand Challenge 2015 drug-like ligand datasets. *Journal of Computer-Aided Molecular Design* **30**, 829–839 (2016).
28. Beckstein, O., Fourrier, A. & Iorga, B. I. Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the OPLS-AA force field. *Journal of Computer-Aided Molecular Design* **28**, 265–276 (2014).
29. Colas, C. & Iorga, B. I. Virtual screening of the SAMPL4 blinded HIV integrase inhibitors dataset. *Journal of Computer-Aided Molecular Design* **28**, 455–462 (2014).
30. Xie, B. & Minh, D. D. L. Alchemical Grid Dock (AlGDock) calculations in the D3R Grand Challenge 3: Binding free energies between flexible ligands and rigid receptors. *Journal of Computer-Aided Molecular Design* (2018).
31. Andrew, R. J. *et al.* Lack of BACE1 S-palmitoylation reduces amyloid burden and mitigates memory deficits in transgenic mouse models of Alzheimer's disease. *Proceedings of the National Academy of Sciences* **114**, E9665–E9674 (2017).
32. Raven, F. *et al.* Soluble Gamma-secretase Modulators Attenuate Alzheimer's β-amyloid Pathology and Induce Conformational Changes in Presenilin 1. *EBioMedicine* **24**, 93–101 (2017).
33. Olson, M. E., Harris, R. S. & Harki, D. A. Apobec Enzymes as Targets for Virus and Cancer Therapy. *Cell Chem Biol* **25**, 36–49 (2018).