PROJECT SUMMARY

Overview:

The University of Colorado, Colorado State University, and the University of Utah propose the formation of a distributed, three-person experienced Cyberteam to provide needed cyberinfrastructure (CI) support to researchers at institutions in the Rocky Mountain Advanced Computing Consortium (RMACC) encompassing the states of Colorado, Idaho, New Mexico, Utah, and Wyoming. Based on experience supporting experienced and new researchers needing CI, it is obvious that there is a new, profound need for "cradle to grave" support for all for the complexity inherent in the reuse of data for reproducibility and extension to new discovery. Researchers need assistance in both publishing their data for further research, and finding, accessing, and using others' published data. Next generation workflows must be supported, while maintaining intellectual property, security, and privacy. This proposal is to add one Cyberteam expert at each of the three lead institutions to provide researcher support, training, and outreach for all RMACC institutions. UNIVERSITY OF COLORADO - User support for CI "cradle to grave" next generation workflows, encompassing data access, preprocessing (including "data carpentry" for reuse), storage, transport, computation, and postprocessing (including basic and advanced scientific visualization, interpretation, and preservation). COLORADO STATE UNIVERSITY - Support for data management and preservation, including making their shared digital repository and data management expertise available to all in the RMACC region. This is particularly germane for smaller institutions, which have no mechanism for this. UNIVERSITY OF UTAH - Maintenance of appropriate security, privacy, and access controls in the entire workflow process. Exploration of usable workflows through "templates" or similar virtualization technologies.

Intellectual Merit :

Formal "cradle to grave" CI workflows will be established, used in training, and published for preservation. Contemporary workflows might include: (a) strategies and methodologies for finding, accessing, and reusing data, (b) preprocessing data from multiple sources and in multiple formats into input files, (c) computation, (d) storage and transport, and (e) postprocessing. Two innovations in data management and curation also are proposed: DSpace will be modified to (a) harvest stored datasets automatically for each research project for reporting purposes, and (b) extend metadata and data standards to store CI workflows, research protocols, electronic lab-books, and research ecosystems. Ten research projects from the three institutions are presented as examples of projects that will be enhanced via this activity. Additional data intensive projects will be engaged through an outreach and engagement program to benefit RMACC institutions with an emphasis to under-resourced organizations.

Broader Impacts :

Broader impacts are the specific target of this effort. (a) In addition to providing CI support for researchers at CU, CSU, and Utah, the proposed Cyberteam will benefit the smaller institutions in the region, who have been allocated 10% of the new NSF-funded HPC system installed at the University of Colorado. Through past activities working with smaller institutions, we have learned that there is in most cases no support for researchers at such institutions. (b) The DSpace shared digital repository at CSU will be expanded for use by the RMACC. CSU has significant expertise in data management, curation, and preservation, and DSpace will accommodate data deposits of all types and sizes from institutions within the region. (c) The Cyberteam support provided via this activity will facilitate a paradigm shift from research in silos to one of data sharing and reuse, from both the provider side (depositing and preserving metadata and datasets for posterity) and consumer side (finding and accessing appropriate datasets that are available from other researchers). This manifests significant culture change involving numerous expert researchers, ensconced in their traditional behaviors. (d) Lessons learned and results will be (1) integrated into courses and curricula at the three lead institutions, (2) used to train students, and (3) broadly disseminated in local, regional, and national venues, including the RMACC annual HPC symposium, Westnet, XSEDE, ACE-Ref, and CaRC. (e) Finally, this proposed Cyberteam will aid several minority-serving institutions within RMACC, and many of the smaller institutions have a high proportion of first-generation and at-risk students.

1 Overview

The Rocky Mountain Advanced Computing Consortium (RMACC), represented by the University of Colorado Boulder, the University of Utah and Colorado State University, propose to create a distributed cyberinfrastructure (CI) team of data and workflow facilitators ("Cyberteam" or "CI Facilitators") for experimental and observational science (EOS). Advances in the number and diversity of data sets require enhanced capabilities to access, reuse, process, analyze, understand, curate, share, and preserve data. A critical aspect in dealing with such large, diverse data sets is to provide expert support for efficient and effective workflows involving data generation, data analysis, visualization, and preservation. Typically, data management, analysis, sharing and curation have been the responsibilities of individual researchers. As a result, data can be difficult to reuse by anyone other than the originators. Likewise, computational and data generation workflows are often cobbled together, hard-coded, and not readily amenable for sharing. These problems must be addressed to make workflows and data available for reuse, and to reuse others' data. The proposed work will fund three FTEs to provide such support for science projects on the three partner campuses as well as projects in the region under the auspices of the RMACC. We note that while each of these institutions having strong support for research computing and the libraries, none at present have facilitators of this type.

This project will strengthen the already existing RMACC culture of collaboration and sharing by bringing together Cyberinfrastructure leadership, both in research computing and in the libraries, to form a regional collaboration that will enhance the manner in which workflows and data management are handled in science and engineering projects. Details of how this proposal addresses recommendations from a recent DOE workshop on "Management, Visualization, and Analysis of Experimental and Observational Data (EOD) - The Convergence of Data and Computing" (Bethel-2016) and the NSF workshop report "The Role of Regional Organizations in Improving Access to the National Computational Infrastructure" (Monaco-2015) are provided below.

2 Project Motivation

2.1 Data Management Support Needs

Although prevalent throughout the research computing enterprise, a lack of support for data curation is particularly pernicious in long-tail science (Heidorn-2008), described as science groups that have a paucity of resources (Heidorn-2008). Such groups are often managed by one lead researcher, with only a part-time commitment to the project, and a small number of graduate or postdoctoral students (Heidorn-2008). These small resources make it difficult, if not impossible, to establish efficient and effective workflows and to support data management. As a result, data from these science groups are often kept offline and out of the public eye – hence the term "dark data" (Heidorn-2008). And, if the data are made public, often the data sets are not properly curated.

To prevent data siloes across projects, institutions that house scientists must provide the infrastructure (both hardware and personnel) to assist these researchers toward a universal goal: high-quality, well-maintained workflows and/or datasets that can be reused and reproduced

(Ogburn-2010). Such efforts benefit scientists by simplifying the process of compliance with funding agency requirements for long-term data archiving to the benefit of all: funding agencies, universities, and the general science community, allowing easy validation of published works and speeding discovery by advancing the starting point through reusing data for new research.

2.2 Data Workflow needs

Challenges with data workflows are manifold: reproducibility, data reuse, and customization or optimization, as often workflows are relatively static and involve complex data processing and dependence on a series of different analysis software. These include homegrown, one-off solutions, complicated software frameworks such as Root, open source community maintained frameworks such as Python or R, commercial software such as Matlab, IDL, SAS, STATA, and Mathematica, and a myriad of evolving workflow toolsets including Pegasus, Radical-Pilot, Saga, Galaxy, Copernicus, Workflow/Workqueue, and Swift. This diversity makes it especially challenging to leverage advances in hardware for improved performance while providing portability across multiple emerging architectures. The challenge for portability resides in simplifying researcher workflows for finding, requesting, and mapping their data and tasks onto the available infrastructure and processing services, while assuring the data always conforms to the required archival standard (such as version control, metadata annotation, etc.), and access control.

Often the challenge for reuse is lack of expertise or knowledge of what data and tools are available, exacerbated by challenges with multiple different and diverse workflows and data sources. This is precisely the area this proposal addresses by fostering a support infrastructure across campuses and the region.

2.3 Protected Data Needs

Further challenges relate to the growth in research computing of restricted and protected data, where higher levels of security, control, auditing, and trust are required. Examples of such data at our institutions are data nascent to new intellectual property, protected health information (PHI), human genomics data (which, although this data is not treated as PHI yet, cannot be de-identified), and export control restricted data such as nuclear engineering. A further challenge is when the compute workflow and data management needs span institutions through collaborations or make use of cloud resources. To handle these usage cases at the University of Utah, over the past seven years we have deployed and maintained a "protected environment" (PE) of HPC (high-performance computing) resources, virtual machines, Windows statistics servers, and associated storage in a more siloed and secure environment (Bradford-2014). See https://www.chpc.utah.edu/resources/ProtectedEnvironment.php. What means are available to properly handle restricted or protected data is an area that is growing rapidly and requires integration into our support environments.

3 The Curation and Data Workflow Facilitators Program

Both the recent XSEDE 16 conference and an NSF CI practitioner workshop explored the need for CI workforce development. This proposal focuses on the *people* aspect of the advanced computing and data ecosystem, which is the current critical need, especially with the increased

dependency of more disciplines on advanced CI. Navigating this complex computing and data environment requires highly trained interdisciplinary experts. The work funded through this proposal will create a connected community of experts in the region that will bridge the gap between scientists and engineers (CI consumers) and local, regional, and national CI providers. Because the principals of this project have been key partners in the RMCMOA (Rocky Mountain Cyberinfrastructure Mentoring and Outreach Alliance) program, we will incorporate the lessons learned from that project and dedicate 20% effort of the CI facilitators to outreach to researchers and educators at smaller institutions. We propose to hire three facilitators with complementary skills and expertise, and integrate them into our local campus and regional efforts. The focus of these facilitators will be as follows:

- CSU: focus on data curation and metadata.
- CU-Boulder: focus on data and compute workflows.
- Utah: focus on protected data and compute workflows

These facilitators, in collaboration with others in the region, will provide and develop regional, shared resources to support data management efforts of research groups, including under-resourced regional partner institutions (such as Utah State University, Colorado School of Mines, University of Wyoming, Idaho State University, University of New Mexico).

3.1 Sustainability

All three institutions are committed to the community building approach of the RMACC. Each campus has committed through the CIO and the Libraries that these positions will be institutionalized and sustained.

3.2 Cyberinfrastructure for Data Workflows and Data Management

In addition to local campus resources, the project team will make the following resources available to researchers in the region with priority for under-resourced campuses and long-tail research groups.

- **Compute for data-intensive workflows:** This capability is provided through the MRI award of the RMACC supercomputer, where 10% of the cycles are already available to smaller institutions in the RMACC region.
- **Data transfers:** Networks in the region, both LAN and WAN, are being upgraded via the RMACC collaboration, and in concert with the RMCMOA grant. The three institutions have modernized their networks, with research science DMZs, Globus, and much higher capacity to support voluminous needs. The partners also have setup data transfer nodes using Globus for efficient transport of large data sets.
- **Storage:** CU-Boulder will provide 100 TB of storage on the active storage service and 200 TB of storage on the archival storage service of the PetaLibary storage service. This system will be used for preservation of the information on the shared DSpace digital repository at CSU, and CSU will supply at least 100 TB of storage. Utah will initially make available at least 200 TB of object storage, in addition to the standard data resources available.
- **Data sharing:** All storage resources for this project are interfaced through data transfer nodes using Globus Online. We will provide Globus sharing that enables researchers to share data among collaborators for all storage resources included in this project.
- **Regional metadata catalog:** Using the OAI-PMH standard, we will create a regional metadata catalog pertinent to the RMACC region.

- **Regional DSpace repository:** CSU has over a decade of experience operating digital repositories, and already operates a shared digital repository, DSpace, for seven libraries in the region. This robust and mature infrastructure will be extended for this project. Researchers in the RMACC region will be able to deposit metadata using Handles and persistent DOIs with their published data, workflows, lab notebooks, research protocols, and research environments. Dataset archival is still very nascent, with a variety of options for the researcher: 1) deposit into a public data repository, e.g., PubMed or other, 2) deposit into a disciplinary repository, 3) deposit into an institutional repository, 4) maintain the dataset where it was generated, or 5) local or regional storage. The approach proposed herein accommodates all of these choices, allowing the researcher to make the choice best suited for the research.
    - Small datasets (currently 20 GBytes or smaller) – Such datasets may be deposited into DSpace for free, at no cost to the user.
    - Medium data sets (greater than 20 GBytes and less than 1 TByte) – Such datasets will be deposited in DSpace and stored at cost, or users may keep these datasets on a local server or other server. We have performed a study of storage costs, and observed that the costs follow the distribution of a Maclaurin series. As this series converges, we will be able to maintain the files indefinitely at a cost of twice the current cost of storage, refreshing storage every five years.
    - Large data sets (greater than 1 TByte) – our recommendation is that a copy of these data sets are stored on the PetaLibrary for long term storage and remain on the system where they are generated for active processing.
- **Publishing of repeatable workflows:** To provide support for repeatable data workflows, we will provide DOIs for virtual machines, containers, and other virtualization approaches that are used to enable repeatable workflows. The project team will also explore how to construct "templates," under the concept developed by the CloudLab at the University of Utah, for cloud and domain science resources to easily prescribe, instantiate, and archive their data life cycle, which includes the analytical workflow but is much broader. The template concept stems from CloudLab's "profile," and is populated when a user creates an experiment, guiding the user to specify all required resources, such as compute, network, storage, and data. The profile is then used to launch the experiment. The profile is as self-complete as possible to enable reproducible experimentation. By extending the CloudLab profile to include necessary descriptions for storage and data archives, and adding the capabilities to invoke external data services, repositories, or workflows (such as preprocessing scripts, campus HPC data transfer nodes (DTN), and file system parameters), we will enable researchers to easily map their data workflows onto CloudLab, HPC infrastructures, and ultimately the national CI.

3.3 Broadening Participation

One of the unique aspects of the RMACC is that it fosters an opportunity for larger schools to collaborate with smaller schools. The NSF funded Summit supercomputer is the first resource that has 10% of its cycles dedicated to under-resourced institutions. This project will augment this resource by providing additional, needed support personnel. We will engage the smaller universities that are a part of RMACC or otherwise local to the three primary schools (CU Boulder, CSU, and University of Utah) in the following meaningful ways:

- At the beginning of Year 1, the PIs from this project will reach out to regional schools to obtain an inventory of research projects requiring data workflow and management support. We will seek one or two research projects from each school that need technical assistance surrounding metadata creation and support, curation, and supporting large workflows on Summit.
- Once the research project(s) are identified, the RMACC Cyberteam will work closely with the researchers and local IT staff to get their project up on the regional Summit HPC system, work through their challenges, and develop a data life cycle plan for their project.
- We have included funds for regional travel for the CI facilitators to visit the smaller schools to meet with their researchers and IT staff. The first engagement will provide the project with a long term data management workflow process that can be used as a template for future projects, a fully curated dataset that is shared on DSpace, training on tools to assist with this effort in the future, and training on data analytics and visualization techniques that researchers can understand and is transferrable to other projects.

3.4 Opportunities for Student Engagement, Education, and Training

Senior personnel Dr. Shelley Knuth, assistant director for research data and training, will lead the engagement, education, and training effort. Educating all researchers, but particularly those in the early stages of their career, will ensure that data are managed properly throughout their life cycle. Such training will empower researchers to utilize the techniques they will learn to apply to their future research long after the life of this project. The training will encompass:

- Workshops offered in person and remotely centered on proper data management, analytics, and visualization techniques.
- Short video modules (less than ten minutes each) on each topic designed to give researchers the basic information needed without an in-depth workshop.
- An informal brown bag lunch, offered weekly, where members of research teams will be invited to get together (in person or remotely) to discuss their current data issues. These lunches will be facilitated by the Cyber team, but are designed to have research groups work together on solving their research problems with high-level assistance from project staff. These type of efforts have proven to be very successful in the region.
- "Train the on-site staff and the trainers," including, data librarians, research data specialists, and other personnel on some of the more technical aspects of data management. Senior personnel Andrew Johnson, CU's data librarian, will be responsible for assisting Knuth in developing training suited to those in the Libraries, and for promoting these trainings among the Library community

3.5 Professional Development for Regional Cyberinfrastructure Professionals

We will organize a yearly workshop as part of the RMACC symposium that will address the professional development needs of regional CI professionals. These sessions will be open to all in the region and will focus on the following topics:

1. Soft Skills: We will provide training to all regional CI professionals on communication with researchers, best practices in training and documentation, defining and discussing reproducible research, and present approaches to construct data sets to facilitate reproducible research. Additionally, training in proposal development and publications will be offered.

2. Data Analytics: Data organization, data cleaning, basic relevant Unix commands, version control, data management with databases, and statistical analysis.
3. Data curation and metadata - Describe relevant Dublin core fields to complete in the "lightweight" metadata collection template. Explore metadata extensions using XML. Hands-on deposit of metadata into DSpace will be the culmination of this training.
4. Data visualization - We already offer such training in our RMACC annual Symposia, but we will strengthen that and also fold in data management and workflows. Topics will include color theory, visualization software, and proper techniques.
5. Workflows - Contemporary workflows also will be presented, discussed and made available publically.

3.6 Exchange Visits to other regional CI organizations

We will provide the opportunity for the facilitators to work at other institutions so that they can gain valuable experiences from different CI providers. This will enhance the expertise of the facilitators and strengthen the collaboration among RMACC CI providers.

3.7 Community Building and Participation in National Cyberinfrastructure Activities

3.7.1 RMACC activities

RMACC system administrators have already been meeting quarterly in person for the last two years. We will organize and sponsor meetings on data curation for the sysadmins, the CI facilitators, and other RMACC CI professionals to create a regional network of CI professionals who can draw on each other's expertise to support solving complex interdisciplinary science problems. Additionally, we will continue to collaborate with Marla Meehl to coordinate our efforts with the NSF funded RMCMOA grant to reach out to small schools and others in the region to educate and advance cyberinfrastructure in the region. Marla also will continue to integrate the efforts of the NSF funded award, "Women in IT Networking at SC (WINS)" within the region to increase and expand diversity. In these activities, she will leverage existing venues including Westnet, The Quilt, and RMACC annual symposium.

3.7.2 XSEDE Campus Champions Program

The Campus Champions program supports campus representatives as a local source of knowledge about high-performance and high-throughput computing and other digital services, opportunities, and resources. This knowledge and assistance empowers campus researchers, educators, and students to advance scientific discovery. RMACC was accepted as a region as part of the regional campus champions. In our recent HPC symposium, we had representation from XSEDE and there was discussion of how to grow regional participation and how to better support under-resourced institutions as part of the regional activities.

3.7.3 ACI-Ref and the CaRC consortium

The University of Utah has been a key partner of the NSF ACI-Ref project from the beginning, and CU-Boulder has committed to participate in the recently awarded Research Network "Advancing Research and Education Through a National Network of Campus Research Computing Infrastructures - The CaRC Consortium." Our effort will coordinate with these

national fora for the exchange and dissemination of best practices, expertise, and technologies, enabling the advancement of campus-based research computing activities.

3.7.4 Big Data Innovation Hub

The PI was recently elected to the steering committee of the West Big Data Innovation Hub. The data hub program will build partnerships across academia, industry, nonprofits, and government to utilize the potential of big data to address societal problems. This Cyberteam project will explore partnership and participation in the Hub as appropriate.

4 Research Activities Supported

In this section, we present ten exemplar funded research projects from all three institutions as non-exclusive examples of projects that will benefit directly from this activity.

4.1 Colorado State University

4.1.1 Jorge Rocca, Electrical and Computer Engineering, ERC for Extreme Ultraviolet Science and Technology

As most advanced electronic circuits and nanoscale machines continue to shrink below the wavelength of visible light, conventional optical technologies are reaching their limits. As a result, light in the Extreme Ultraviolet (EUV) region of the spectrum (wavelengths of approximately 3 to 50 nm) is poised to become a key element in technologies. Our project consists of the development and application of compact EUV lasers with the objective of making EUV technology widely available to solve challenging scientific and industrial problems. The group of Prof. Rocca in the Engineering Research Center (ERC) for Extreme Ultraviolet Science and Technology (EUV ERC) is exploring the development (Reagan-2014) and applications (Kuznetsoy-2015) of compact EUV and soft X-ray lasers with the objective of making EUV technology widely available to solve challenging scientific and industrial problems. The research also involves the study of plasmas that are intense emitters of EUV and X-ray radiation (Avaria-2015, Kaymak-2016, Yin-2016). These data files contain results of numerical hydrodynamic/atomic model simulations, and range from 10 MB to 5 GB in size, with CCD images of the order of 2-4 MB each. The number of files can be several hundred per simulation and up to several hundred per day can be generated. Such data must be produced, shared, curated and preserved. **NSF Funding: IIP-1343456 $800,000, ECCS-1509925 $400,000; DOE Funding: DE-0000000SC14610 $830,000, DE-FG02-04ER15592 $1,774,998, DE-SC0016136 $1,155,000; DOD-USAF Funding: FA9550-14-1-0232 $899,950.**

4.1.2 David Randall, Atmospheric Science

Randall's research group hosts the NSF Science and Technology Center for Multiscale Modeling of Atmospheric Processes (CMMAP) that performs very high-resolution global atmospheric modeling, with a particular focus on the role of clouds in climate change (Benedict-2009, Benedict-2011, DeMott-2011, Stan-2010, Thayer-Calder-2009, Randall-2015, Kooperman-2016, Randall-2016a, Randall-2016b). Some models use ~10^{10} grid cells, necessary to represent individual large clouds in the global atmosphere, and are run on a variety of supercomputers. The models routinely produce multiple TB of data in a single run, and produce close to a PB of

model output. The data are being analyzed by scientists from various institutions across the United States, with the goal of understanding the role of clouds in climate change. **NSF Grants: ATM-0425247, $37,505,835; AGS-1461270, $595,766, AGS-1500187 $656,746; DOE Grant: DE-SC0008226, $459,978; NOAA Grant: NA13OAR4310103, $196,000; LLNL Grant: B614354, $452,781.**

4.1.3 Christos Papadopoulos, Computer Science, LANDER IMPACT, (Information Marketplace for Policy and Analysis of Cyber-risk & Trust)

Via a centralized repository, this project provides developers and evaluators with regularly updated network operations data relevant to cyber defense – the traces so collected that are made available to researchers involved in security research. The traces include continuous packet header and flow captures, DDoS attacks (Hussain-2016), spam logs, black lists, and periodic (every three months) censuses and surveys for the entire Internet (Heidemann-2008). CSU's role as a Data Host requires that periodically we curate data for others. For example, we curate a 7 TByte trace from Packet Clearing House (PCH) containing a DARPA attack dataset. **DHS Funding: NBCHC080035, $3,000,000**

Supporting Scientific Applications over NDN. In this project, we are building a prototype system to serve scientific data (Climate and High-energy Physics) over Named Data Networking (NDN). The data is served over the CSU Science DMZ network through an NDN testbed overlay in ESnet and is used by scientists at LBNL, Caltech and South Korea. Our system can cache about 200TB of data, that we use to test the performance and characteristics of our prototype. Or plan is to expand our system to host the majority of the popular data in these domains [FaS15]. **NSF Funding: 13410999, $1,000,000**

4.1.4 Dennis Ojima and Jeff Morisette, Natural Resource Ecology Lab, and Department of Interior North Central Climate Science Center

The Secretary of the Interior, Secretary Kenneth Salazar, announced the intent to create the DOI North Central Climate Science Center (NC CSC) on October 21, 2011. The mission of the NC CSC is to provide the best available climate science and synthesis to inform energy, land, and cultural resource management within the North Central Domain. These activities include linking climate data with ecosystem response models (Glick-2011), which will require large amounts of both input and output data and the transfer of those data among the eight other university consortium member, federal agencies, and non-governmental organizations. The NC CSC coordinates research among the nine universities in its consortium and over 20 USGS research offices, as well as other federal, state, and NGO collaborators. The center needs to curate large amounts of data (on the order of TBs) for the joint research involving the various integrated models to facilitate analysis and discovery. **USGS Cooperative Agreement Funding: G11AC90009, $2.2M/yr.**

4.1.5 Wanllenstein, Ross, Stargell, Yao, and VandeWoude, Bioinformatics Institute

With the decreasing cost of generating massive amounts of "omics" data, there comes new challenges in data analysis that require access to infrastructure and expertise in HPC.

Wallenstein's work is focused on the characterization of soil microorganism response to changes in plant productivity and climate through a variety of metagenomic, proteomic, and metabolomics approaches (Ernakovich-2015, Osborne-2015). Ross is working on the development and optimization of a structure-based algorithm for the prediction of prion propensity. Stargell's research is focused on investigation of the activation of poised RNA polymerase using yeast and the response to oxidative stress as the regulatory system (Chen-2016, Kuo-2015). This project relies heavily on ChIP-seq and RNA-seq and numerous large datasets are produced and require curation. Yao is focused on identifying behavior of chromatin-associated proteins (Long-2014a, Long-2014b). VandeWoude is evaluating how the interaction between landscape structure and management interventions affects disease spread in populations of wide-ranging apex predators (Puma) (Bevins-2012). Their integrative approach will utilize large scale next generation sequencing and geospatial tools to advance models that will predict and mitigate virulent disease outbreaks for wide-ranging species in complex landscapes which will require produce numerous, large datasets requiring curation. **NSF Funding: PLR-1255228 $968,906, 1517231 $595,510, MCB-1023771 $748,000, MCB-1330019 $500,000, MCB-1158323 $830,000; DOE Funding DE-0000000SC10568, $940,966; NIH Funding: 5R01GM105991-02 $1,500,000, 1R01GM098401-01 $1,500,000.**

4.2 University of Colorado Boulder

4.2.1 Cognitive and Neuroscience

This project involves the Cognitive and Affective Neuroscience Lab in the Department of Psychology and Neuroscience at CU-Boulder. This group routinely captures brain activity through magnetic resonance imaging (MRI) to understand the pathways that generate and regulate pain and emotion, cognitive performance, lifespan development, and psychopathology. Datasets are several hundred terabytes, and are currently being stored on the PetaLibrary. Efforts by their group parallel and are integrated with efforts nation- and world-wide to aggregate large databases of neuroimaging data to enable advanced analyses that will revolutionize our understanding of the mind, brain, mental health, and neurological disorders. Their group in particular, supported in part by a multi-site collaborative R01 and "big data" supplement grant (R01DA035484 and 3RO1 DA0353484), is focused on building an affective neuroscience database containing neuroimaging data from many sites around the world, and providing tools that will make this data accessible over the web for download and in-browser machine learning analysis. For this project to succeed, gold-standard curation of the database is essential.

The Wager lab has been proactive in their data management efforts – the group has recently hired a data management expert, but a lack of data curation expertise in this group has thus far prevented these curation practices from taking place. The data management expert hired as part of the Wager group is in a perfect position to gain valuable data curation expertise when working with a CI facilitator to ensure their data collection is properly managed. This will dramatically enhance the group's efforts to make available a large-scale database of neuroimaging data and web-enabled data access and machine learning tools.

4.2.2 Earth Science

Earth Lab is designed to accelerate discovery and generate insights about the pace and pattern of global environmental change through the integration of Earth observations from space-, aero-, and ground-based platforms. This initiative will reduce societal risk and surprise by integrating data and facilitating collaboration networks to better understand and predict slow, fast, and abrupt earth system change. As such, the Earth Lab will link several existing data streams from across the Earth system to one location, with benefit from combined and aggregated information (e.g., PRISM climate data, (Daly-2002). The Earth Lab's data needs are two-fold: first, science experts will analyze data from several data streams including climate metrics and social behavior data (e.g., insurance data, agricultural land use and management data, and economic market data) important to understanding societal response to climate change. To accomplish this goal, proper data management, integration, analytics, and visualization tools must be in place to fully understand and gain knowledge from the data. Second, in the process of analyzing the data, the research scientists will establish a workflow that helps them to answer their research questions, and enables future reproducibility of scientific conclusions. Earth Lab has also taken a proactive approach to ensuring their data needs have been met by hiring data visualization and analytics experts to support the scientists with their data related efforts.

4.2.3 Computational Biology

Sam Flaxman: Recent advances in DNA sequencing technology have opened doors to studying population genetics and evolution at the scale of the whole genome. The Flaxman group has simulated biologically realistic populations and genomes in order to begin developing a novel adaptive framework for interpreting laboratory data (Feder-2014, Flaxman-2013, Flaxman-2014). This will help answer one of the longest standing questions in evolutionary biology: What are the main drivers of speciation? With faster computational resources they are nearing the point of being able to generate powerful and testable predictions about the temporal dynamics of speciation in specific empirical systems, by performing simulations that track thousands of neutral sites in the genome.

Matt Keller: Research by the Keller group helps bridge the gap between psychology and bioinformatics. Understanding the "genetic architecture" of traits—the frequencies, numbers, and effects of genetic variants that cause interpersonal differences—has been one of the central goals of statistical, molecular and evolutionary genetics over the last fifty years. In the midst of this deluge of data, how- ever, fundamental questions about the genetic architecture of traits remain unanswered or are poorly characterized (Keller-2014). They are working to develop numerically-intensive methodologies that will greatly reduce the uncertainty surrounding the genetic architecture of traits.

4.3 University of Utah

4.3.1 Ensembles of molecular simulations

Thomas Cheatham: Using large ensembles of molecular dynamics simulations on XSEDE and Blue Waters, the Cheatham lab has demonstrated the ability to reproducibly converge the conformational ensembles of DNA duplexes, RNA tetranucleotides, and RNA tetraloops (Galindo-Murillo-2016, Cheatham-2015, Galindo-Murillo-2014). This provides a means to

validate, assess and improve the underlying methods and "force fields" and also provides insight into nucleic acid structure, dynamics and function. Large ensembles generate considerable data and a challenge is not only the compute workflow but extensive data management. Although some of our data is available on the WWW (http://amber.utah.edu) and our "scripts" are published in supporting data, this project is far from true curation and share-ability of the workflow and data. **NSF ACI-1521728, ACI-1443054, ACI-1341034, ACI-1341935 and CHE-1266307 and NIH R01-GM098102.**

4.3.2 Genomics workflow and data management pipelines

UStar Center for Genetic Discovery (UCGD) and the HCI Bioinformatics and Sequencing Core: At the HCI core facility, there are a variety of homegrown pipelines for sequencing, annotating, meta-genomic, and other processing of both open and restricted genomics data. Examples are Tomato, Gnomex, Opal/VASST, and others. A facilitator in CHPC that could help further educate on restricted data processing and management, and also on alternative compute workflow and data management tools would further enable their research and operations. **UCGD was founded based on a $6,000,000 gift from the USTAR initiative and the University of Utah and is supported by numerous NSF grants (IOS-1126998, IOS-1561337) and NIH grants (UM1-HL128711, U01-HL131698, R01-GM099939, R01-GM104390, R01-HG008628, U01-HG006513, and R01-HG006693) and the HCI core supports many funded researchers across campus.**

5 Expertise of the Project Team

PI Hauser is the Director of Research Computing at CU-Boulder, has decades of experience using and developing and deploying research cyberinfrastructure. He is the organizer and lead of RMACC. He will be responsible for the overall project execution and coordinating the efforts with the Summit supercomputer project.

Co-PI Burns is the VP for IT and Dean of Libraries at CSU, and has forty years of experience as a CI researcher and educator. He is responsible for research computing, central IT, and DSpace at CSU. Together with Williams, he will be responsible for DSpace enhancements, and provide expert librarians to participate with the CI facilitators.

Co-PI Cheatham is Professor of Medicinal Chemistry and Director of Research Computing and the Center for High Performance Computing at the University of Utah. He has considerable experience with the national CI including large computer time allocations and volunteer duties with both XSEDE and Blue Waters. In addition, he is Co-PI of the ACI-REF consortium of institutions sharing CI facilitators. His research area is biomolecular simulation, analysis and workflows and his lab is one of the core AMBER software development teams. He will help mentor the facilitators with a particular focus on environments for protected data and repeatable workflows.

Co-PI Siegel – Abell Endowed Chair Distinguished Professor of Electrical and Computer Engineering and Professor of Computer Science at CSU, is well versed in a variety of areas of CI research and education. As the founding director of the university-wide CSU Information Science and Technology Center (ISTeC: istec.colostate.edu) from 2002 to 2013, he has extensive

experience in engaging faculty in collaborative activities. As part of the governance structure of this new NSF effort, he will coordinate with his counterparts at the various other institutions to ensure we are providing coherent, consistent cyberinfrastructure services and support for research and education at all of the member institutions.

Co-PI Williams, Dean of CU Libraries, has decades of experience as a library leader, participates in data management activities, and is an expert in the area of metadata. Together with Burns, he will be responsible for oversight of DSpace enhancements and for outreach to and integration with associated regional and national efforts.

Senior Personnel Knuth, Assistant Director of Research Data and Training in CU's Research Computing, will work on the educational training portion of this project, developing the curricula for, managing, and teaching the tutorials offered. She is also the director of Earth Lab's Analytics Hub, which, among other things, facilitates and guides collaborative efforts between data experts and the scientists. Knuth will facilitate efforts between the CI facilitators and her Analytics Hub staff as the two groups work closely on Earth Lab's science projects to enhance and deepen their knowledge of data workflow processes.

Senior Personnel Johnson, the CU-Boulder Data Librarian, will work with the CU science projects and the CI facilitators to establish and support proper metadata and curation practices. He will also use his expertise and those of the CI facilitators to establish a best practices data management model for units on campus that can serve as a template for other research groups.

6 Goals and Milestones

Project tasks and timelines are presented below. A three-year effort is proposed. Those responsible for the tasks are shown in parentheses.

1. Governance (PI and co-PIs) – implement governance structure (Quarter 1, Year 1), hold regular meetings, prepare and conduct marketing and communications activities, collect feedback from participants and redirect activities as appropriate. Meetings will occur as often as required to produce excellent results, and at least quarterly.
2. Hire Cyberteam members (PI and co-PIs): This will be done collaboratively across the three institutions to ensure that activities are coordinated and complementary. This activity is expected to take about six months to get the new CI facilitators on board. Will be completed within Quarters 1 and 2 of Year 1.
3. Engagement with the science teams (Cheatham, Hauser, Siegel). This activity will occur continuously throughout the grant period.
    a. Understand existing workflows for exemplar and additional selected projects.
    b. Develop new, optimized workflows for exemplar and additional selected projects.
    c. Implement security and privacy provisions locally and regionally - four months of effort.
4. Cyberinfrastructure Development - the activities below will proceed sequentially, because of staff availability.
    a. DSpace instance (Burns) – deploy, test and verify an RMACC DSpace instance for the participants. Share new code in Github using a Creative Commons license.

i. DSpace access– Explore and implement authentication and authorization methods for the RMACC participants - two months of effort. (Quarter 1, Year 1)
ii. DSpace self-deposit – Modify the DSpace self-deposit module to support this activity, including a requirement for "lightweight" metadata pertinent to specific grants for later harvesting - two months of effort. (Quarter 2, Year 1)
iii. DSpace metadata harvesting – A web interface will be developed for harvesting metadata associated to individual grants for reporting compliance with Data Management Plans - two months of effort. (Quarter 3, Year 1)
iv. DSpace data type extensions - The traditional metadata and data frameworks will be extended to accommodate virtual machines, containers, and other virtualization approaches that are used to enable repeatable workflows - four months of effort. (Quarter 2 of Year 2)

b. PetaLibrary access for researchers outside of CU-Boulder (Hauser) - two months of effort, and adding users as needed over the grant period. (Quarter 4 of Year 1)

5. Exchange of metadata with other digital repositories in the RMACC (Williams and Burns). Here we will use the OAI-PMH standard for harvesting metadata between and among the other digital repositories in the region, to develop a comprehensive view of project activities - two months of effort. (Quarter 1 Year 3)
6. Training (Knuth) – Training sessions will be delivered face to face on campus, at the RMACC annual HPC Symposia, and other venues as appropriate, including the Westnet affinity group covering the RMACC region. (Beginning in Quarter 1 of Year 2 and henceforth ongoing)
7. Review of activities of the program (Hauser). During each years RMACC symposium the project team will organize a retreat to evaluate the effectiveness of the program. We will make data driven decision on how to improve the cyberinfrastructure and the engagement efforts by using surveys and interviews with the science teams.
8. Ensuring proper knowledge transfer to appropriate librarians at CU and other institutions (Johnson) - ongoing.
9. Reporting to the NSF (PI) – The PI will prepare annual reports and the final report in concert with the co-PIs. (ongoing)
10. Dissemination of results (PI and co-PIs) – Will include lectures in classes, dissemination of results at conferences, including the Westnet Conference and EDUCAUSE. (ongoing)

7 Management Plan

A strong, well-coordinated project structure will ensure that the project is maximally successful. Fortunately, such a structure is already well established under the auspices of the RMACC. Experienced and knowledgeable personnel are already engaged and will be responsible for managing all aspects of this project, including internal and external relationships, relevant systems, the implementation, communications, training, and coordination. We will use the following committees to direct the work of the facilitators:

- Management committee (existing under the RMACC): Hauser, Cheatham, Siegel
    - Will make decisions on projects selected for support from RMACC institutions outside the campuses
    - Oversight of the project

    - Participate in training and education
    - Responsible for communications and reporting
- Data and technical advisory committee (addition of responsibilities to the existing DSpace Policy Committee): Burns and Williams
    - Provides the technical lead on metadata approaches, and technical architecture
    - Provides oversight of the DSpace development projects
    - Advise on integration of this effort with national efforts, e.g. the Digital Public Library of America (DPLA), ARL's SHARE initiative, and other national efforts

8 Broader Impact

Numerous broader impacts for the proposed infrastructure upgrade have been identified. Among them are:

- This activity will provide CI support and a repository (DSpace) for all institutions in the RMACC region, especially important for the smaller institutions.
- This activity will also distinctly benefit existing DSpace participants via the development of projects described in Section 6, notably including self-deposit, researcher dataset harvesting for reporting, and expansion of metadata and data types to support research workflows.
- The activity will be included in Computer Science and Computer Information Systems classes at the three institutions.
- The infrastructure will serve as a critical element in transforming the culture of our faculty and graduate students about referencing, reusing, depositing, curating, and preserving data, and will serve as a model for other funding agencies, which are expected to follow NSF in requiring data management.
- Research results will be presented, referenced, deposited (as appropriate), and then made accessible, reliable, and reusable, enhancing the speed and scope of discovery with positive effects on scientific and economic development.
- The enhanced activity will support and facilitate our mutual efforts to share expertise, resources, strategies and techniques with our RMACC partners. Especially critical in this activity are collaborations among faculty and researchers that are enabled by the support.
- The culture of data management will become institutionalized as a result of this proposal, as we believe is the intent of the NSF.
- Finally, this project will serve several minority-serving institutions, and many of the smaller institutions have a high proportion of first-generation and at-risk students.

9 Results of Prior NSF Support

Hauser, Siegel, and Burns received **"MRI Collaborative Consortium: Acquisition of a shared supercomputer by the Rocky Mountain Advanced Computing Consortium"** ACI-1532235 $2.7 million (09/01/15-08/31/18) to procure, install, operate, and manage a shared HPC system under the auspices of the RMACC. The system purchased under these awards is currently being installed and undergoing acceptance testing at the University of Colorado. Intellectual merit: the system will serve a broad array of researchers at the two institutions. Broader impact: The University of Colorado will operate and support it on behalf of the RMACC (10% of cycles are dedicated to the smaller RMACC institutions). No publications have been produced yet.

MRI: Acquisition of a Scalable Petascale Storage Infrastructure for Data-Collections and Data-Intensive Discovery, ACI-1126839, Hauser, Gallaher, Williams, Banich, Guralnick, $699,945, 09/10/2011 – 08/31/2015. Intellectual merit: The PetaLibary storage system consists of two storage pools: active, mounted on the RC condo cluster, and archive, which is an HSM system. The PetaLibrary active storage is used by more than 20 research groups or departments, with over 500 TB reserved and 400 TB in active use. Some examples of research projects being supported include the storage, sharing, and analysis of fMRI data by the Institute of Cognitive Science; statistical study of genomics data by the Institute for Behavioral Genetics; and storage of experimental data by several biochemistry groups. The archive storage of the PetaLibary is currently used by 15 research groups, who have reserved over 400 TB and are actively using 450 TB. A major focus for Archive users is the storage of digital collections: both the CU Libraries and CU Museum of Natural History have substantial digitization efforts underway that require large secure storage resources. Other projects taking advantage of the PetaLibrary archive include groups in psychology/neuroscience and chemical/biological engineering who are storing experimental results to comply with data management requirements. Broader impact: To support the data needs of the CU-Boulder community, the recommendations of a data management task force report (Rankine-2012) were implemented. The CU-Boulder Libraries and Research Computing have established a joint data management service, a data faculty advisory committee, and a data management executive committee. Additionally, the Libraries and Research Computing conducted numerous data-management training sessions to train the next generation of scientist and engineers in best practices (Johnson-2014).

Co-PI Cheatham has an active research group developing and applying large-scale biomolecular simulation approaches in the AMBER software on campus, XSEDE and Blue Waters resources. **PI:ACI-1515572 and OCI-1036208: "PRAC – Ensembles of molecular dynamics engines for assessing force fields, conformational change, and free energies of proteins and nucleic acids"** is a travel award with a substantial allocation of computer time (12M node hours per year, $40,000, 9/1/15-7/31/18) on the Blue Waters Petascale Resource. Intellectual merit: Demonstrating convergence and reproducibility in elucidating conformational ensembles. Broader impact and products: Assessment and validation of force fields and release of GPU optimized AMBER 14, 15, and16 with M-REMD support. A number of publications have resulted so far (Bergonzo-2015, Bergonzo-2015, Bergonzo-Iii-2015, Cheatham-2015, Dissanayake-2015, Galindo-Murillo-2015, Hopkins-2015, Lai-2015, Maier-2015, Nguyen-2015, Robertson-2015, Thibault-2015, Zgarbova-2015, Bergonzo-2016, Galindo-Murillo-2016, Galindo-Murillo-2016, Waters-2016).

Bibliography

G. Avaria, M. Grisham, J. Li, F.G. Tomasel, V.N. Shlyaptsev, M. Busquet, M. Woolston, J.J. Rocca, "Extreme Degree of Ionization in Homogenous Micro-Capillary Plasma Columns Heated by Ultrafast Current Pulses", Physical Review Letters **114**, 095001 (2015).

Bevins SN, Carver S, Boydston EE, Lyren LM, Alldredge M, Logan KA, et al. Three pathogens in sympatric populations of pumas, bobcats, and domestic cats: implications for infectious disease transmission. PLoS One. 2012;7:e31403.

J. J. Benedict and D. A. Randall, "Structure of the Madden–Julian Oscillation in the superparameterized CAM," *Journal of the Atmospheric Sciences*, Vol. 66, No. 11, Nov. 2009, pp. 3277-3296.

J. J. Benedict and D. A. Randall, "Impacts of idealized air-sea coupling on Madden-Julian Oscillation structure in the super-parameterized CAM," *Journal of the Atmospheric Sciences*, Vol. 68, No. 9, Sep. 2011, pp. 1990-2008.

Bergonzo, C., K. B. Hall and T. E. Cheatham, 3rd (2015). "Stem-Loop V of Varkud Satellite RNA Exhibits Characteristics of the Mg(2+) Bound Structure in the Presence of Monovalent Ions." J Phys Chem B **119**(38): 12355-12364.

Bergonzo, C., K. B. Hall and T. E. Cheatham, 3rd (2016). "Divalent Ion Dependent Conformational Changes in an RNA Stem-Loop Observed by Molecular Dynamics." J Chem Theory Comput **12**(7): 3382-3389.

Bergonzo, C., N. A. Henriksen, D. R. Roe and T. E. Cheatham, III (2015). "Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields." RNA: [in press].

Bergonzo, C. and T. E. Iii (2015). "Improved Force Field Parameters Lead to a Better Description of RNA Structure." J Chem Theory Comput **11**(9): 3969-3972.

Bethel, E.W. ed., 2016. *Management, Visualization, and Analysis of Experimental and Observational Data (EOD) The Convergence of Data and Computing Workshop Final Report*, Office of Advanced Scientific Computing Research, DOE Office of Science Bethesda, Maryland.

Bradford, W., J.F. Hurdle, B. LaSalle, and J.C. Facelli, "Development of a HIPAA-compliant environment for translational research data and analytics." *J Am Med Inform Assoc*, 2014. 21(1): p. 185-9.

TE Cheatham, III and DR Roe. "The impact of heterogeneous computing on workflows for biomolecular simulation and analysis." *Computing in Science and Engineering* 17:2, 30-39 (2015).

Daly, C. et al., 2002. A knowledge-based approach to the statistical mapping of climate. *Climate Research*, 22(2), pp.99–113.

Dissanayake, T., J. M. Swails, M. E. Harris, A. E. Roitberg and D. M. York (2015). "Interpretation of pH-activity profiles for acid-base catalysis from molecular simulations." Biochemistry **54**(6): 1307-1313.

Chen X, D'Arcy S, Radebaugh CA, Krzizike DD, Giebler HA, Huang L, et al. Histone Chaperone Nap1 Is a Major Regulator of Histone H2A-H2B Dynamics at the Inducible GAL Locus. Mol Cell Biol. 2016;36:1287-96.

C. A. DeMott, C. Stan, D. A. Randall, J. L. Kinter III, and M. Khairoutdinov, "The Asian Monsoon in the Super-Parameterized CCSM and its relation to tropical wave activity," *Journal of Climate*, Vol. 24, No. 19, Oct. 2011, pp. 5134-5156.

Ernakovich JG, Wallenstein MD. Permafrost microbial community traits and functional diversity indicate low activity at in situ thaw temperatures. Soil Biology and Biochemistry. 2015;87:78-89.

Fan, C., Shannigrahi, S., DiBenedetto, S., Olschanowsky, C., Papadopoulos, C. and Newman, H. Managing scientific data with named data networking. In Proceedings of the Fifth International Workshop on Network-Aware Data Management (co-located with Supercomputing 2015), Austin, TX, November 2015.

J. L. Feder, P. Nosil, S. M. Flaxman, *Front. Genet.* 5,
http://dx.doi.org/10.3389/fgene.2014.00295
(Aug. 2014).

S. M. Flaxman, J. L. Feder, P. Nosil, *EVOLUTION* 67, 2577–2591 (Feb. 2013).

S. M. Flaxman, A. C. Wacholder, J. L. Feder, P. Nosil, *Mol Ecol* 23, 4074–4088 (May 2014).

R Galindo-Murillo, DR Roe, and TE Cheatham, III. "On the absence of intrahelical DNA dynamics on the μs to ms timescale." *Nature Commun.* 5:5152 (2014) doi: 10.1038/ncomms6152

Galindo-Murillo, R., J. C. Garcia-Ramos, L. Ruiz-Azuara, T. E. Cheatham, 3rd and F. Cortes-Guzman (2015). "Intercalation processes of copper complexes in DNA." Nucleic Acids Res.

R Galindo-Murillo, JC Robertson, M Zgarbová, J Šponer, M Otyepka, P Jurečka, and TE Cheatham III. "Assessing the current state of Amber force field modifications for DNA." *J. Chem. Theory Comp.* (2016) DOI: 10.1021/acs.jctc.6b00186 [in press]

Galindo-Murillo, R., D. R. Davis and T. E. Cheatham, 3rd (2016). "Probing the influence of hypermodified residues within the tRNA3(Lys) anticodon stem loop interacting with the A-loop primer sequence from HIV-1." Biochim Biophys Acta **1860**(3): 607-617.

P. Glick, B.A. Stein, and N.A. Edelson, *Scanning the Conservation Horizon: A Guide to Climate Change Vulnerability Assessment*, National Wildlife Federation, Washington, D.C, 2011.

Heidorn, P.B., 2008. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2), pp.280–299.

J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister, "Census and survey of the visible internet," *ACM Internet Measurement Conference*, Oct. 2008, pp. 169-182.

Hopkins, C. W., S. Le Grand, R. C. Walker and A. E. Roitberg (2015). "Long time step molecular dynamics through hydrogen mass repartitioning." J. Chem. Theory Comp. **11**: 1864-1874.

A. Hussain, J. Heidemann, and C. Papadopoulos, "Identification of repeated denial of service attacks," *IEEE Infocom 2006*, Apr. 2006.

A. Johnson, S. Knuth, Best Practices for Good Data Management, accessed on 02/18/2015 (2014; http://researchcomputing.github.io/Boot_Camps/pdfs/2014_fall_best_practices_data_management.pdf).

V. Kaymak, A. Pukhov, V.N. Shlyaptsev, J.J. Rocca, "Nanoscale ultradense Z-pinch formation from laser irradiated nanowire arrays", Physics Review Letters, **117**, 035004, (2016).

J. Gratten, N. R. Wray, M. C. Keller, P. M. Visscher, *Nature Neuroscience* 17, 782–790 (May 2014).

Kooperman, G. J., M. S. Pritchard, M. A. Burt, M. D. Branson, and D. A. Randall, 2016: Robust effects of cloud super-parameterization on simulated daily rainfall intensity statistics across multiple versions of the Community Earth System Model. J. Adv. Modeling Earth Syst., 8, doi:10.1002/ 2015MS000574.

I. Kuznetsov, J. Filevich, F. Dong, M. Woolston, W.L. Chao, E.H. Anderson, E.R. Bernstein, D.C. Crick, J.J. Rocca, C.S. Menoni, "Three-dimensional nanoscale molecular imaging by extreme ultraviolet laser ablation mass spectrometry", Nature Communications **6**, 6944 (2015).

Kuo YM, Henry RA, Huang L, Chen X, Stargell LA, Andrews AJ. Utilizing targeted mass spectrometry to demonstrate Asf1-dependent increases in residue specificity for Rtt109-Vps75 mediated histone acetylation. PLoS One. 2015;10:e0118516.

Lai, C. T., H. J. Li, W. Yu, S. Shah, G. R. Bommineni, V. Perrone, M. Garcia-Diaz, P. J. Tonge and C. Simmerling (2015). "Rational Modulation of the Induced-Fit Conformational Change for Slow-Onset Inhibition in Mycobacterium tuberculosis InhA." Biochemistry **54**(30): 4683-4691.

Long L, Furgason M, Yao T. Generation of nonhydrolyzable ubiquitin-histone mimics. Methods. 2014;70:134-8.

Long L, Thelen JP, Furgason M, Haj-Yahya M, Brik A, Cheng D, et al. The U4/U6 recycling factor SART3 has histone chaperone activity and associates with USP15 to regulate H2B deubiquitination. J Biol Chem. 2014;289:8916-30.

Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling (2015). "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB." J Chem Theory Comput **11**(8): 3696-3713.

G. E. Monaco, D. F. McMullen, G. Huntoon, J. Leasure, D. Swanson, H. Neeman, Jo. Blake, and K. Adams, The Role of Regional Organizations in Improving Access to the National Computational Infrastructure, https://drive.google.com/file/d/0B9RBtxud9RbBemdYcGpnOExPUmM/view?usp=sharing, 2015.

Nguyen, H., A. Perez, S. Bermeo and C. Simmerling (2015). "Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins." J. Chem. Theory Comp. **11**: 3714-3728.

Ogburn, J.L., 2010. The Imperative for Data Curation. *portal: Libraries and the Academy*, 10(2), pp.241–246.

Osborne B, Baron J, Wallenstein MD. Moisture and temperature controls on nitrification differ among ammonia oxidizer communities from three alpine soil habitats. Frontiers of Earth Science. 2015;10:1-12.

Randall, D. A., 2015: An Introduction to the Global Circulation of the Atmosphere. Princeton University Press, 442 pp., ISBN-13: 978-0691148960.

Randall, D. A., C. DeMott, C. Stan, M. Khairoutdinov, J. Benedict, R. McCrary, and K. Thayer-Calder, 2016: Simulations of the tropical general circulation with a multiscale global model. Chapter 15 of Multiscale Convection-Coupled Systems in the Tropics: A tribute to Dr. Michio Yanai, R. G. Fovelland W.-W. Tung, Eds. Meteorological Monographs, 56, published by the American Meteorological Society. DOI: http://dx.doi.org/10.1175/AMSMONOGRAPHS-D-15-0016.1.

Randall, D. A., A. D. Del Genio, L. J. Donner, W. D. Collins, and W. A. Klein, 2016: The Impact of ARM on Climate Modeling. Chapter 26 of The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years, D. D. Turner and R. Ellingson, Eds. Meteorological Monographs, 57, published by the American Meteorological Society. DOI: http://dx.doi.org/10.1175/AMSMONOGRAPHS-D-15-0016.1.

P. Rankin, R. Duerr, T. Hauser, A. Johnson, J. Maness, M. Parsons, H. Rajaram, R. Shoemaker, K. Stacey, A. Viggio, J. Wakimoto, “Research Data Management at the University of Colorado Boulder”, tech. rep. (University of Colorado Boulder, 2012), http://hdl.handle.net/10971/1398.

B.A. Reagan, M. Berrill, K.A. Wernsing, C. Baumgarten, M. Woolston, J.J. Rocca, "High-average-power, 100-Hz-repetition-rate, tabletop soft-x-ray lasers at sub-15-nm wavelengths," Physical Review Applied **89**, 053820 (2014).

Robertson, J. C. and T. E. Cheatham, 3rd (2015). "DNA Backbone BI/BII Distribution and Dynamics in E2 Protein-Bound Environment Determined by Molecular Dynamics Simulations." J Phys Chem B **119**(44): 14111-14119.

C. Stan, M. Khairoutdinov, C. A. DeMott, V. Krishnamurthy, D. M. Straus, D. A. Randall, J. L. Kinter, III, and J. Shukla, "An ocean-atmosphere climate simulation with an embedded cloud resolving model," *Geophysical Research Letters,* Vol. 37, Jan. 2010.

K. Thayer-Calder and D. A. Randall, "The role of convective moistening in the formation and progression of the Madden–Julian Oscillation," *Journal of the Atmospheric Sciences*, Vol. 66, No. 11, Nov. 2009, pp. 3297-3312.

Thibault, J., D. R. Roe, K. Eilbeck, T. E. Cheatham, III and J. C. Facelli (2015). "Development of an informatics infrastructure for data exchange of biomolecular simulations: Architecture, data models, and ontology." SAR and QSAR in Environ. Res.

Handle System, Wikipedia, last edit July 2016.

Waters, J. T., X. J. Lu, R. Galindo-Murillo, J. C. Gumbart, H. D. Kim, T. E. Cheatham, 3rd and S. C. Harvey (2016). "Transitions of Double-Stranded DNA Between the A- and B-Forms." J Phys Chem B.

L. Yin, H. Wang, B.A. Reagan, C. Baumgarten, E. Gullikson, V. N. Shlyaptsev, M. Berrill, J.J. Rocca, '6.7-nm emission from Gd and Tb plasmas over a broad range of parameters for beyond-extreme-ultraviolet lithography," Physical Review Applied, in press (2016).

Zgarbova, M., J. Sponer, M. Otyepka, T. E. Cheatham, 3rd, R. Galindo-Murillo and P. Jurecka (2015). "Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA." J Chem Theory Comput **11**(12): 5723-5736.