Rutgers Discovery Informatics Institute: Accelerating discovery in the State of New Jersey through high performance computing
Rutgers Discovery Informatics Institute: Accelerating discovery in the State of New Jersey through high performance computing
TABLE OF CONTENTS

1  A Message from the Director of RDI
2  The Road to Caliburn
4  Architecting Caliburn
8  Introducing RDI²
9  Science Impact
39  Outreach and Development Efforts
40  RDI² Staff
42  Key Words/Topics

SELECTED ABSTRACTS

10  Examining the Structure of Proteins that Contribute to Neurodegenerative Diseases
12  Galaxy Evolution Studies Enabled by Large Galaxy Surveys
14  Ab initio QM/MM Development in the Amber Molecular Dynamics Program
16  Computational Investigation of Coke Formation on Platinum Alloy Catalysts
18  Massively Parallel Optimization Methods in Machine Learning and Decision Analysis
20  Catalytic Strategies of RNA Enzymes
24  High Fidelity Simulations of Shock Wave Reflection Inside a Shock Tube
26  Large Eddy Simulation of Shock Wave Turbulent Boundary Layers Interaction
28  Deep Learning of Protein-protein Interactions and Molecular Dynamics Simulations of the Orai Calcium Channel
30  Computational Study of Molecular Adsorption on MnO₂ nanoscaffolds
32  Direct Numerical Simulation of Blood Flow in Physiologically Realistic Microvascular Networks
34  Large Scale Simulations of Proteins and Nucleic Acids
36  Using Parallel HPC to Determine Ramsey-Theoretic Quantities
DEAR COLLEAGUES,

Computing has become an integral part of science and society. As a result, to be competitive it is essential that researchers, students and industry have ready access to state-of-the-art computing resources, expertise and training. In fact, it is now widely accepted that, to out-compete, one must out-compute!

It is this recognition that has driven the Rutgers Discovery Informatics Institute (RDF) to take on a leadership role in ensuring that all researchers and students at Rutgers, and across New Jersey, have free and open access to computing resources that are at par with leading resources worldwide, ensuring that they are nationally and internationally competitive. Over the past years, RDF has provided strategic and tactical leadership in this space. Its efforts have included the deployment and operation of the Excalibur supercomputing system in partnership with IBM, leadership in strategic planning for research computing at Rutgers, and the establishment of the Rutgers Office for Advanced Research Computing (OARC).

The acquisition, deployment and operation of the Caliburn system (including the modular datacenter that houses it) represents a culmination of these efforts. The Caliburn journey started with the conceptualization of the vision for an advanced research computing ecosystem at Rutgers and beyond, translating this concept into a fundable proposal, and acquiring funding from the State through the New Jersey Higher Education Equipment Leasing Fund (ELF) to support its execution. Caliburn was uniquely architected to leverage the state-of-the-art technologies at the time (e.g., the interconnect and the memory subsystem), to meet the current and envisioned needs of the targeted research community, and to position Rutgers and New Jersey as a leader in research computing nationally. With over 23,000 cores, Caliburn can perform over 800 trillion floating point operations per second. When it was commissioned in Summer 2016, Caliburn was ranked on the Top500 list of computer systems worldwide as the #2 system among the US Big 10 institutions and #8 among all US academic institutions, #50 among academic institutions globally, and #166 among all computer systems worldwide. It was also universally lauded for its balanced architecture and use of innovative technologies.

To date, Caliburn has delivered hundreds of millions of computing hours to researchers and students across New Jersey, driving research and innovation in all areas of science, engineering, business and medicine, and enabling research and discoveries that would not be possible otherwise. This report highlights some of these research results from projects enabled by Caliburn.

RDF fundamentally changed the research computing landscape at Rutgers and for the state, and established models and resources that are on par with leading institutions nationally and internationally. I do hope that others will build on the foundation that RDF has established and continue to invest in providing leading-edge computing capabilities to researchers and students at Rutgers and throughout New Jersey.

Sincerely,

Manish Parashar, IEEE Fellow, AAAS Fellow, ACM DS
Founding Director, Rutgers Discovery Informatics Institute (RDF)
Distinguished Professor, Dept. of Computer Science
The Road to Caliburn

ADVANCED RESEARCH COMPUTING ESSENTIAL TO SCIENCE AND SOCIETY

Advanced research computing is playing increasingly critical roles in driving research and innovation in all areas of science, medicine, engineering, and business. As a result, to be competitive it is essential that researchers, students and industry have ready access to state-of-the-art computing resources, expertise and training, and it is also essential that academic institutions provide this access.

STRATEGIC PLANNING FOR RESEARCH COMPUTING AT RUTGERS

In 2013, Rutgers Discovery Informatics Institute (RDI2) led a Rutgers-wide strategic planning process for research computing, which resulted in the June 2014 report *Accelerating Innovation Through Advanced Cyberinfrastructure: A Strategic Vision for Research Cyberinfrastructure at Rutgers.* The report was a result of community driven efforts including multiple systematic surveys, numerous focus group meetings with faculty, researcher groups, systems staff and administrators, and round-table discussions. Key findings included the critical importance of a comprehensive state-of-the-art research computing ecosystem at Rutgers, the lack of existing research computing leadership and resources, need for a bold vision for research computing and an execution plan to accomplish this vision. Key recommendations of the report included the need to establish an Office for Research Cyberinfrastructure at Rutgers and to deploy a balanced nationally competitive Advanced Cyberinfrastructure at Rutgers.

ENVISIONING A LEADERSHIP-CLASS RESEARCH COMPUTING CAPABILITY

Over the past years, RDF, in close collaboration with the Rutgers Office of Economic Development, has worked to systematically address these (and other) recommendations, including leading the establishment of the Rutgers Office of Advanced Research Computing (OARC) and conceptualizing, deploying and operating a world-class advanced research cyberinfrastructure at Rutgers. The acquisition, deployment and operation of the Caliburn system (including the modular datacenter that houses it) represents a culmination of these efforts.

The journey started with developing the vision for a research computing ecosystem at Rutgers and beyond, translating this concept into a fundable proposal, and acquiring funding from the State through the New Jersey Higher Education Equipment Leasing Fund (ELF) to support its execution. Specifically, the proposal envisioned “a statewide resource that will have far-reaching benefits to Rutgers and to the entire state; impacting higher education institutions, industry and state government” and a system that “will be seamlessly accessible as a cloud service, providing researchers, students, industry and government across the state” and be “the most powerful academic system in the state.”

Caliburn was uniquely architected to leverage the state-of-the-art technologies at the time, to meet current and envisioned needs of the research community by balancing large-scale computations with big data analytics capability, and to position Rutgers as a leader in research computing nationally. With over 23,000 cores, Caliburn can perform over 800 trillion floating point operations per second. When it was commissioned in Summer 2016, Caliburn was ranked on the Top500 list of computer systems worldwide as the #2 system among the US Big Ten institutions and #8 among all US academic institutions, #50 among academic institutions globally, and #166 among all computer systems worldwide.

Caliburn has been universally lauded for its balanced architecture and use of innovative technologies. The RDF team worked closely with the selected vendor (determined by Rutgers procurement process) to develop an architecture that most effectively leveraged innovative, state-of-the-art technologies in computing, storage and networking, and to tune its deployment and operations. For example,

- Caliburn’s architecture was designed to provide the latest innovations in processor and network technology.
- It delivers a significant amount of NVMe (non-volatile memory express) storage which allows faster access to data.
- Caliburn was among the first systems to use the Intel Omni-Path fabric and to equip its compute nodes with NVMe devices – such an architecture is now considered the state-of-the-art.
- Caliburn also provides large scale parallelism that allows researchers to run massively parallel applications with multiple components of the application running concurrently.

This design made Caliburn an asset with unique architecture and capabilities for the research community, while also enabling general purpose computational and data-intensive applications.

**Caliburn: Driving Science, Delivering Innovation**

Caliburn enables researchers to address some of the most complex problems in science and engineering, which a standard computer cannot handle or may take many years to solve. To date, Caliburn has delivered hundreds of millions of computing hours to researchers and students across New Jersey, driving research and innovation in all areas of science, engineering, business and medicine, and enabling research and discoveries that would not be possible otherwise. Users span a wide range of fields including chemistry and chemical biology, engineering, genomics, humanities, integrative biology, mathematics, medical informatics, microbiology, proteomics, physics and astronomy to name some. Grand challenges in these fields cannot be addressed at scale without an instrument such as Caliburn. For example, a research group from Rutgers–Camden is leveraging Caliburn to execute molecular dynamics simulation at unprecedented scales to study diseases such as epilepsy, addiction, schizophrenia, bipolar, and unipolar depression.

RDF’s leadership and the Caliburn ecosystem have fundamentally changed the landscape of advanced research computing at Rutgers and beyond, and provides a solid foundation to build on for the future.

---

To date, Caliburn has delivered hundreds of millions of computing hours to researchers and students across New Jersey, driving research and innovation in all areas of science, engineering, business and medicine, and enabling research and discoveries that would not be possible otherwise.
Building a supercomputer is not about putting together a few fashionable technologies within budget. Building a supercomputer entails combining the right hardware and software in such a way that not only do they work in sync with the best performance possible, but also that the supercomputer becomes a valuable asset for the research community. At the end of the day, the ultimate goal is to accelerate scientific discoveries across multiple fields of research, regardless of the technologies used along the way. Therefore, we knew we had to design something innovative from the ground up, for which we took a data-driven and user-centric approach strategically projected in three phases.

**PHASE I**

The Phase I system, initially collocated in a temporary facility, provided over 200 teraflops of computational and data analytics capabilities, and one petabyte of high-performance storage to faculty and staff researchers throughout the research community. Users of this system spanned a wide range of disciplines including chemistry and chemical biology, engineering, genomics, humanities, integrative biology, mathematics, medical informatics, microbiology, physics and astronomy, and proteomics. Phase I consisted of 144 compute nodes, each of which was equipped with two 12-core Intel Xeon E5-2680v3 processors and 256 GB of main memory. Additionally, 16 of these compute nodes were also furnished with an NVIDIA Tesla K40m graphics processing unit (GPU) accelerator.

The server form factor for the 128 non-GPU nodes is based on Dell PowerEdge M1000e blade enclosures, each one being fully populated with sixteen half-height PowerEdge M630 blade servers whose technical specifications are detailed in Table 1. The compute nodes equipped with GPUs are rack-mounted Dell PowerEdge R730 servers whose technical specifications are shown in Table 2.

Regarding the storage capabilities, the high-performance parallel file system is based on DataDirect Networks (DDN) GRIDScaler, an implementation of IBM Spectrum Scale, previously known as General Parallel File System (GPFS), which provides a global name space, shared file system access, simultaneous file access from multiple nodes, high recoverability and data availability through replication, the ability to make changes while a file system is mounted, and simplified administration even in large environments. The DDN GRIDScaler configuration deployed, whose different components are listed in Table 3, provides 1 petabyte of usable space over both InfiniBand and TCP/IP networks.

---

**Table 1. PowerEdge M630 compute node specifications**

<table>
<thead>
<tr>
<th>Processor</th>
<th>Dual Intel Xeon E5-2680 v3 2.5GHz, 12-Core, 120W processors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>256 GB DDR4-2133 (4 nodes provide 768 GB RAM DDR-2133)</td>
</tr>
<tr>
<td>Storage</td>
<td>Single 300G 10K RPM SAS hard drive</td>
</tr>
<tr>
<td>Network</td>
<td>Mellanox FDR 56 Gbps InfiniBand adapter</td>
</tr>
</tbody>
</table>
Table 2. PowerEdge
R730 compute/GPU node specifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Dual Intel Xeon E5-2680 v3 2.5 GHz, 12-Core, 120W processors</td>
</tr>
<tr>
<td>Memory</td>
<td>256 GB DDR4-2133</td>
</tr>
<tr>
<td>Storage</td>
<td>Single 300G 10K RPM SAS hard drive</td>
</tr>
<tr>
<td>Network</td>
<td>Mellanox FDR 56 Gbps InfiniBand adapter</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA GPU Tesla K40m</td>
</tr>
</tbody>
</table>

Table 3. DataDirect Networks
GRIDScaler configuration

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rack(s)</td>
<td>APC 42U rack</td>
</tr>
<tr>
<td>Controller(s)</td>
<td>SFA12KXE Embedded Application Couple</td>
</tr>
<tr>
<td>Enclosure(s)</td>
<td>5 x DDN S58460 84-Slot 6Gb/s SAS/SATA/SSD Enclosures</td>
</tr>
<tr>
<td>Disks</td>
<td>240 x 6TB 6Gb/s SAS Drives, 8 x 400GB Mixed Use SSD 6Gb/s</td>
</tr>
<tr>
<td>Network</td>
<td>16 x QSFP FDR-InfiniBand ports</td>
</tr>
</tbody>
</table>

In terms of networking, Phase I is based on Mellanox FDR 56Gbps InfiniBand configured in a two-tier fat tree topology. In this configuration, compute nodes are connected to edge switches, and these edge switches are in turn connected to core switches. Four root switches and ten edge switches conform the InfiniBand network fabric.

Overall, this phase brought online 3,456 cores, 38 terabytes of main memory, and 1 petabyte of storage, providing 138 teraflops of peak performance, boosted up to approximately 206 teraflops when taking into account the 68 teraflops of peak performance provided by the GPU server pool.

PHASE II

*Phase II delivered a new self-contained modular data center (MDC) that hosts seventeen 42U racks and delivers 350 kilowatts total power.*

The MDC cooling system is based on chilled water and rear-door heat exchangers. Two coolant distribution units provide closed-loop controlled cooling water to the rear door heat exchangers, requiring controlled secondary water, with a total maximum nominal cooling capacity of 305kW each. The process side water loop, referred to as the “Secondary Circuit”, is a sealed pressurized system where the heat extracted from the rear door heat exchangers is rejected to a chilled water circuit. This chilled water supply loop is referred to as the “Primary Circuit”. The chilled water is supplied from a dedicated chiller skid. The rear door heat exchanger is a water-cooled door that is mounted on the rear of the rack to cool the air that is heated and exhausted by the devices inside the rack. Supply and return water hoses deliver conditioned water to the rear door heat exchanger and remove the heated water from it back to the coolant distribution unit.
PHASE III, INFRASTRUCTURE OPERATIONS, AND USER OPERATIONS

On the subject of power protection, the MDC is backed up by a top of the line Eaton uninterruptible power system (UPS) which provides industry-leading efficiency and scalable battery run-times in a small footprint. Its N+1 redundancy design provides high reliability and eliminates potential for single-point of failure. The model deployed was designed with numerous technological advancements that create an unprecedented level of reliability while emphasizing serviceability, lowering MTTR (mean-time-to-repair) and enhancing availability.

With regard to safety, the MDC uses mechanisms to control climate and maintain appropriate operating temperature and humidity. The MDC is also equipped with an automated fire detection and suppression system based on a clean agent system that removes heat and breaks up the fire at the molecular level, and because it discharges as a gas, it leaves no residue and does not require costly clean-up like water or dry chemical systems.

PHASE III

Phase III encompassed the final installation of the supercomputer and the last elements of the network, which provided over 600 teraflops of computational and data analytics capabilities and approximately 200 terabytes of non-volatile memory express (NVMe) storage, interconnected with an Intel Omni-Path 100Gbps network fabric. This phase comprised 560 compute servers, each containing two 18-core Intel Xeon E5-2695v4 processors, 256 GB of main memory, 400 GB of non-volatile memory, and an Intel Omni-Path Host-Fabric Interface Adapter as detailed in Table 4. The server form factor is based on the Supermicro FatTwin enclosure, which is highly efficient by design. Due to its shared components, the FatTwin improves cost-effectiveness and reliability, while its modular architecture makes it flexible to configure and easy to maintain. 70 4U enclosures are evenly spread across 10 racks, each enclosure containing 8 half-width servers.

Phase III is based on Intel Omni-Path (OPA), a new network interconnect developed by Intel, and it was among the first clusters to use it. The data throughput capability of Intel OPA in its first generation is 100 Gbps in each direction of the link. In addition to high bandwidth, Intel OPA is a low latency and highly resilient interconnect with many different Quality of Service (QoS) features. Caliburn implements a two-tier Omni-Path fabric. Eight core switches and nineteen edge switches conform the Intel OPA network fabric.

Modern parallel file systems such as Lustre or IBM Spectrum Scale are designed to provide high, scalable I/O bandwidth in response to growing I/O requirements; however, the bursty I/O characteristics of many data-

<table>
<thead>
<tr>
<th>Table 4. FatTwin compute node specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>Memory</td>
</tr>
<tr>
<td>Storage</td>
</tr>
<tr>
<td>Network</td>
</tr>
</tbody>
</table>
intensive scientific applications make it difficult for back-end parallel file systems to efficiently handle I/O requests. A burst buffer system, through which data can be temporarily buffered via high-performance storage mediums, allows for gradual flushing of data to back-end file systems. This was precisely the main reason why we decided to set up the compute nodes with local non-volatile memory express (NVMe) drives.

This phase incorporated the bulk of the processing power of the system by adding in 20,160 cores, 140 terabytes of memory, and 218 terabytes of non-volatile memory, with a measured performance of 603 teraflops and a theoretical peak performance of 677 teraflops.

After completing Phase III, collectively, 23,616 cores, 178 terabytes of main memory, 218 terabytes of non-volatile memory, and 1 petabyte of high-performance storage were available to students and researchers at Rutgers University and at other institutions in the state of New Jersey, as well as to industry statewide.

INFRASTRUCTURE OPERATIONS
How effectively a data center is managed is directly dependent upon the effectiveness of the infrastructure management platform. Accordingly, the MDC infrastructure and the supercomputer equipment were configured as a facility protected under the Rutgers Discovery Informatics Institute’s Advanced Monitoring Platform, a secure, robust and reliable solution based on the combination of modern open-source software and in-house developments, which had proven itself as a highly effective monitoring and alerting system for other large scale projects such as the National Science Foundation Ocean Observatories Initiative (OOI) geographically distributed cyberinfrastructure.

It is important to point out that keeping a strict preventive maintenance schedule for each one of the different components in the MDC has been crucial to avoid unexpected downtime.

USER OPERATIONS
We also determined that the cluster not only had to be built based on the community needs but had to be adaptable for almost every researcher throughout the cluster’s lifespan, consequently, we consider user proposed changes, coming in via our ticketing system, as high priority service requests.

Booking time on Caliburn is possible via two different allocation opportunities, startup allocations and awarded allocations, each with distinct eligibility rules. Whilst startup allocations are resource-wise small and can be requested at any time, getting an awarded allocation requires submitting a project proposal that gets reviewed by an expert panel based on criteria such as project motivation, computational effectiveness, availability of resources, etc. Allocations are awarded on regular intervals throughout the year.
We are in an era of Big Data and Extreme Computing! Computing and data have the potential for fundamentally transforming science and society, impacting every aspect of our lives and our environment. They are critical to understanding (and managing) natural, engineering and human systems, from climate change to smart infrastructure, personalized medicine/healthcare and social networks. Consequently, it has become essential for an academic institution to have access to leading edge computing and data capabilities, and more importantly, the multidisciplinary research structures and expertise necessary to effectively leverage them to address grand challenges.

We founded the Rutgers Discovery Informatics Institute (RDI²) to respond to this reality. RDI², a unit of the Office of Research and Economic Development (ORED) at Rutgers University, is a university-wide, multidisciplinary resource for computational and data-enabled science and engineering research, education and outreach. The overarching goal of RDI² is to bridge more traditional research boundaries and catalyze socio-technical changes across all disciplines, stimulating new thinking and new practices essential to addressing and solving grand challenges. RDI² dynamically integrates research, education and advanced technologies to broaden academic and industry access to state-of-the-art computing and data analytics capabilities, provides academic and industry researchers with the interdisciplinary expertise and resources necessary to address the grand challenges in their fields, and provides leadership in Big Data and Advanced Cyberinfrastructure at the university, state and national levels.

RDI² has a strong set of achievements and impacts since its creation in 2012. Combining leadership in research, expertise, advanced tools and algorithms, and advanced cyberinfrastructure ecosystems, RDI² has engaged leading academic and industry researchers (at Rutgers, New Jersey, the US, and internationally) in innovative, interdisciplinary collaborations. These collaborations have resulted in over 170 peer reviewed scholarly publications, multiple awards, and widely used software systems. RDI² PIs have also successfully received 52 federal, state and industry grants and sponsorships for a total of ~$52.5M. RDI² researchers are contributing to large scale national and international projects in areas such as extreme-scale computing, large-scale and streaming data-analytics, edge computing and IoT, and computational and data-driven science and engineering and advanced cyberinfrastructure more broadly.

RDI² has deployed a comprehensive cyberinfrastructure ecosystem at Rutgers, which includes research and production facilities, high performance computing, cloud and data infrastructure, research instruments, and experimental platforms. RDI², through a grant it received from the state of New Jersey, architected, deployed and operates New Jersey’s most powerful supercomputer, Caliburn (introduced in June 2016 as the 2nd most powerful computer among the Big 10). RDI²’s research cyberinfrastructure is open to all faculty and students across Rutgers, as well as academia and industry throughout the state, and is part of the XSEDE National Cyberinfrastructure. Caliburn is delivering hundreds of millions of hours of massively parallel computing to researcher across New Jersey and enabling cutting-edge research and innovations in a wide range of disciplines including chemistry and chemical biology, engineering, genomics, humanities, integrative biology, mathematics, medical informatics, microbiology, proteomics, physics and astronomy, that would not be possible otherwise. Recent innovation have added hardware and software support for machine learning and data analytics as well as seamless Cloud integration.

---

RDI²’s overarching mission is to accelerate discovery and drive innovation through advanced computing and data, and we will continue to push the boundaries of multidisciplinary research and break barriers towards realizing this mission. We invite you to join us in our journey.

– Manish Parashar
RDI is also operating the cyberinfrastructure for the NSF Ocean Observatories Initiative (OOI) program, the largest ocean observing program in the world. RDI, in collaboration with academic partners including Pennsylvania State University, and regional network providers, KINBER and NJEdge, designed the Virtual Data Collaboratory (VDC) to transform shared data as a modality for research and discovery. The VDC is an NSF-funded regional federated data cyberinfrastructure for data-intensive, interdisciplinary and collaborative research.

RDI also led the university-wide ACI strategy that resulted in the formation of the Rutgers Office of Advanced Research Computing (OARC) and was instrumental in the formation of the statewide New Jersey Big Data Alliance (NJBDA). The NJBDA brings together academic institutions, government organizations and industry across the state to collaboratively address Big Data challenges and seize Big Data opportunities. RDI is also playing a leadership role in the state’s cyberinfrastructure and Big Data efforts.

RDI has been actively involved in education and outreach activities in the areas of computation and data at the graduate, undergraduate, and K-12 levels, has established a fellowship program to support students engaged in computational and data-enabled science and engineering research, and has been working with the community to explore social issues such as cyberbullying and its impact on students. RDI’s distinguished seminar series regularly hosts leading researchers from academia, governments and industry. RDI also organizes workshops and other events on important and timely topics.

According to Manish Parashar, Founding Director of RDI and Distinguished Professor of Computer Science, “RDI’s overarching mission is to accelerate discovery and drive innovation through advanced computing and data, and we will continue to push the boundaries of multidisciplinary research and break barriers towards realizing this mission. We invite you to join us in our journey.”
Examining the Structure of Proteins that Contribute to Neurodegenerative Diseases

PI: Grace Brannigan; Student: Ruchi Lohia, PhD Candidate; Institution: Rutgers University–Camden, Center for Computational and Integrative Biology and Department of Physics

EXECUTIVE SUMMARY:

The toll that neurodegenerative diseases can take on a person and their loved ones can be daunting. At the very least, the uncertainty about how to care for a person and to offer compassionate treatment and care is hampered by a lack of understanding about the disease’s causes. Much research has been done to identify the biochemistry of the brain, and to discover how the delicate interplay of neurons and chemical signals play a role in giving life to the mind.

One of the pathways for understanding how the mind works is to study how the brain’s biochemistry is affected by mutations of proteins that are known to be important in the growth of neurons and intracellular signaling. The studies conducted by our research group seek to find out why certain gene mutations are tied to aging and stress related disorders, as well as lower tolerance for drugs. To do this, molecular dynamics simulations are an indispensable tool to understand how the mutations affect the structure and dynamics of proteins. Since large proteins are modeled atom by atom and require long simulation times, their dynamical behavior has to be followed using a large number of compute cores in order to run the computational experiments efficiently and get meaningful results.

RESEARCH CHALLENGE:

An important mental function that can be a precursor or symptom of neurodegeneration is memory impairment. The signaling protein, brain derived neurotrophic factor (BDNF), is particularly noted for its importance in regulating neural development through the growth of neural synapses and intracellular signaling.

Val66Met single nucleotide polymorphism (SNP) is one of the earliest identified variants in the prodomain region BDNF. An extensive library of genome-wide association (and even earlier) studies have repeatedly identified the Val66Met SNP as reducing hippocampal volume and episodic memory, as well as predicting increased susceptibility to neuropsychiatric disorders including schizophrenia, bipolar, and unipolar depression, but associations have been inconsistent and population dependent. (Figure 1)

Our research attempts to elucidate the underlying molecular mechanism via which the Val66Met affects conformation of prodomain region of BDNF and eventually increases susceptibility to neuropsychiatric disorders.

METHODS AND CODES:

We carried out 128µs of fully atomistic explicit solvent molecular dynamics (MD) simulations with temperature replica exchange (64replicas, 2µs per replica) of the 90-residues BDNF prodomain with and without Val66Met substitution to investigate its conformational effects.

T-REMD: is a widely used replica exchange enhanced sampling technique where multiple replicas of the system are run at different temperatures with the goal of improving sampling by making the conformations at higher temperature available to the lower ones and vice versa. At specific intervals during the simulations, the conformations of the two replicas are swapped according to a Metropolis-type criterion.

Gromacs 5.1.2-mpi simulations package (has already been installed on Caliburn).

RESULTS:

Two simulations with 64 parallel replicas each were run on Caliburn. Since the simulation convergence time increases exponentially with increased protein length, our simulations ran for approximately 0.5 milliseconds,

Figure 1. a) Cartoon representation of proBDNF and it’s two domains, prodomain and BDNF. Prodomain is an IDP and it contains the Val66Met SNP. b) Computationally generated model of prodomain. The Val66 is shown in green color. We explore the possible mechanism by which Val66Met SNP can affect the conformation and dynamics of the BDNF protein.
which is among some of the longest simulation time for a disordered protein. These long simulations successfully reproduced the earlier experimentally observed (NMR) variables from Anastasia et al. 2013. (Figure 2)

We find that the prodomain BDNF, can be meaningfully divided into domains based on sequence alone. We identified two unique hydrophobic domains, the SNP containing domain and the Janus domain. These unique domains have biological significance as well: both SNP domain and Janus domain are equally essential for intracellular trafficking of proBDNF.

We observe that, only M66 sequence interacts strongly with the only other Methionine in its sequence due to preferred Met-Met interactions. This points out that although both Valine and Methionine residues are hydrophobic, they do not essentially have the same interaction preference. Met-Met interactions have been under-appreciated in IDP’s. (Figure 3)

The prodomain BDNF receptor SorCS2 also has the presence of several Methionines’s on its surface. Therefore, our finding suggests that M66 binds strongly to this receptor due to Met-Met interactions and not Val66, this leads to differential biological function for Val66 and Met66 sequence.

**ROLE OF CALIBURN AND RDI:**

Caliburn’s nodes, each containing 18 cores, are perfectly suited for providing parallelism. With partitions that allow numerous simultaneous node usage, researchers can take advantage of running single massively parallel simulations as well as ensemble workflows.

**REFERENCES:**


**PUBLICATIONS AND DATA SETS:**

Galaxy Evolution Studies Enabled by Large Galaxy Surveys

**PI:** Eric Gawiser,
**Students:** Humna Awan and Kartheik Iyer, PhD Candidates
**Institution:** Rutgers University–New Brunswick, Department of Physics and Astronomy

**EXECUTIVE SUMMARY:**
Understanding the physical processes that drive galaxy evolution is one of the primary challenges of 21st century astrophysics. Observations of distant galaxies using large telescopes enable astronomers to peer back in time to the early stages of the universe. While space-based telescopes like the Hubble Space Telescope (optical frequencies), the Spitzer Space Telescope (infrared frequencies) and Chandra Satellite (X-ray frequencies) allow taking high-resolution images of deep space without being hindered by the atmosphere, the ground-based facilities like the Hobby-Eberly Dark Energy Experiment (optical frequencies) and the upcoming Large Synoptic Survey Telescope (optical frequencies) allow observing vast parts of the sky for extragalactic science. These telescopes probe complementary parts of the electromagnetic spectrum and their observations allow us to understand how galaxies evolve with time and how they form stars. Studying these characteristics of galaxies as a function of time by reconstructing their star formation histories, and as a function of space by studying how they cluster at different spatial scales holds immense power as we are able to probe larger questions like how our Milky Way galaxy came to be and why our universe is accelerating.

In order to address these questions, robust statistical analysis of the large datasets (millions to tens of billions of galaxy observations over the next ten years) must be carried out in order to study the evolution of galaxies. Since we aim to apply new statistical tools on large datasets, containing $10^5$ to $10^7$ galaxies, both the development and the application of statistical tools requires high performance computing resources such as those offered by Caliburn.

**RESEARCH CHALLENGE:**
In order to study galaxy evolution, distances to distant galaxies must be estimated accurately. This is done by measuring the galaxies’ redshift – a quantity that relates the frequency of the observed light to how far the galaxies are, possible due to the finite speed of light and our knowledge of the intrinsic frequency of the emitted light. Hence, using multiwavelength observations of the galaxies collected with a combination of ground- and space-based telescopes, we are able to estimate galaxy redshifts. This process is rather challenging; not only do we have incomplete datasets for the different types of galaxies, the currently available tools do not account for measurement uncertainties and hence are unable to fully harness the power of large telescopes. Our team tackles these challenges on two fronts: 1) by developing methods that extract valuable features of galaxies as a function of time, hence allowing us to probe galaxy evolution more robustly and infer their distances, and 2) by developing tools that account for uncertainties in inferred distances for low-variance measurements of galaxy clustering as a function of time.

Specifically, we develop statistical methods to extract information about a galaxy’s star formation history (SFH) from their spectral energy distributions (SEDs), which are measurements of how bright a galaxy is across a range of wavelengths. We extract these SEDs from $O(10^5$-$10^9)$ galaxies from current surveys as well as those from mock catalogs of the upcoming surveys, and since individual galaxies can be run through the developed pipeline independently, this process can be efficiently parallelized. However, in order to infer SFHs for different galaxies, SED fitting must be performed while accounting for galaxies at similar redshifts – the step that cannot be fully parallelized as the fitting of different galaxies must communicate which each other.

We also develop tools for quantifying galaxy clustering, using 2-point correlation functions which are naively $O(N^2)$ operations and $O(N\log N)$ at best (e.g., using TreeCorr$^*$). The new estimators for the correlation functions account for redshift uncertainties directly, which entails optimizations on galaxy catalogs of size $O(10^5$-$10^9)$ – a computational challenge which must be undertaken to be able to utilize the power of large datasets.

**METHODS AND CODES:**
In order to carry out the SED fitting on large datasets, we built pre-grids to do the fitting on galaxies in the same redshift bins; loading these grids into shared memory therefore allows for a fast fitting algorithm. Then, we perform posterior probability distribution calculations for individual galaxies using a large array of goodness-of-fit values, and run tests and quality checks of noisy telescope data. Using these overarching ideas, we built the Dense Basis SED fitting method that reconstructs the SFHs of galaxies using a large atlas of physically-motivated SFHs assembled from various functional forms adopted in the current literature and their combinations. The technique is effectively non-parametric and uses statistical techniques to tune the number of SFH components based on the amount of information in the SEDs being fit. Recent developments to the Dense Basis method entail using the technique of Gaussian Processes to improve computational efficiency and achieve independence from predetermined functional forms.
As for the improved clustering estimators, we have developed a methodology to account for probabilistic distance measurements when measuring galaxy clustering. By weighting each galaxy with its probability of being at a certain distance, we are able to measure galaxy clustering without the need of forward modelling. Furthermore, we are able to optimize the weights by considering numerous galaxy samples; this optimization allows for a one-step measurement of the correlation functions, while providing a platform to account for various dependencies and prior information directly into the estimator.

RESULTS:

The Dense Basis method has been scaled up from a sample of ~1000 galaxies (Iyer & Gawiser, 2017) to ~15,000 galaxies (Iyer et al., 2018) by parallelizing it to run on Caliburn. After algorithmic improvements including improvements using gaussian processes, the method has been run to extract SFHs from nearly ~ 60,000 galaxies (Iyer et al., 2019) and is currently being used to perform tests on ~200,000 galaxies, each being fit 100 times as a part of a calibration study. The extracted SFHs have been used to study how galaxies evolve on average at different epochs of the universe and at different stellar masses (figure below; left panel). Furthermore, Iyer et al. (2018) performed a novel analysis by propagating galaxies backwards in time along their SFHs to study their properties when they were younger and less massive (figure below; right). Observationally, this would require ~11 times more telescope time to perform the same analysis, since lower mass galaxies are much fainter and require longer exposures to study directly, so our ability to measure these properties as a function of time breaks new ground.

Also, our new estimator is scaled to be applicable to tens of galaxy samples, each containing ~200,000 galaxies; this is achieved by parallelizing the various runs and using the large memory resources offered by Caliburn and Elf. The details of the new methodology as well as its validation on mock galaxy catalogs are presented in Awan & Gawiser, 2019 (in prep).

ROLE OF CALIBURN AND RDI²:

RDI²’s cyber infrastructure allows for a cost effective and scalable framework to perform compute-intensive calculations needed for scientific discovery in the Big Data era of astronomy; development of these methods would be impossible without the help of the responsive staff and the hardware provided by the center.

PUBLICATIONS AND DATA SETS:


¹ https://github.com/rmjarvis/TreeCorr
Ab initio QM/MM Development in the Amber Molecular Dynamics Program

**PI:** Timothy Giese  
**Institution:** Rutgers University–New Brunswick, Laboratory for Biomolecular Simulation Research

**EXECUTIVE SUMMARY:**
Advances in computer hardware have made practical the routine application of *ab initio* quantum mechanical (QM) methods within molecular dynamics (MD) simulations of small molecules, ligands, and ribozyme active sites on a timescale that is long enough to extract the thermodynamic quantities necessary to make a meaningful comparison with experimental measurements. The integration of QM methods within MD software has traditionally focused primarily on the use of semiempirical Hamiltonians, whereas *ab initio* methods were of limited use due to the small amount of sampling that could be afforded. As a consequence, MD programs “loosely-coupled” their software with existing *ab initio* QM packages in a manner that did not require changes to the QM software. That is, the MD program would write an input file for the QM program at each step of dynamics, evaluate the QM energy via a “system call” to the QM executable, and then read the energy and forces from the QM software’s output file. This strategy places severe limitations on the interaction between the QM and molecular mechanical (MM) portions of the system. We developed an *ab initio* QM library that can be directly linked into Amber’s MD executable to facilitate the development of the QM/MM interactions. Using this strategy we implemented a particle mesh Ewald (PME) method for periodic *ab initio* QM/MM electrostatics, and we showed that the use of electrostatic cutoffs – as used in the loose-coupling strategy – leads to nonphysical simulation artifacts, such as large reaction barriers and artificial solvation shells near the cutoff boundary. Now that we have the ability to perform periodic, condensed-phase *ab initio* QM/MM simulations, we are using Caliburn to push the limits of what can be achieved through its use in various chemical applications. The remainder of this report discusses an example application that we have completed, but has yet to be published.

**RESEARCH CHALLENGE:**
The twister ribozymes are a class of nucleolytic ribozymes that have recently been structurally characterized. Conflicting interpretations of different crystal structures, and in some cases conflicting interpretations of the same functional data has led to debate about its architecture and mechanism. We have attempted to use a comprehensive computational RNA enzymology approach aimed at providing a unified interpretation of existing structural and functional data. We use GPU-accelerated free energy methods with enhanced sampling to ascertain microscopic nucleobase pKa values of the implicated general acid and base, from which predicted activity-pH profiles can be compared directly with experiments. We have performed *ab initio* QM/MM simulations with full dynamic solvation under periodic boundary conditions to determine mechanistic pathways through multidimensional free energy landscapes for the reaction. We have also characterized the rate-controlling transition state, and made predictions about kinetic isotope effects and linear free energy relations.

**METHODS AND CODES:**
We performed the *ab initio* QM/MM simulations with a development version of the SANDER program within the AMBER software package. The solvated twister ribozyme system contains 66k atoms. The active site contains 59 QM atoms, and the QM energy is evaluated with the PBE0/6-31G* hybrid density functional method, polarized by the MM charges within 12 Å of any QM atom and the long-range, periodic PME interactions corresponding to all atoms outside the 12 Å cutoff. Kinetic isotope effect values (KIE) are obtained by geometry optimization for a transition state structure and evaluating the Hessian, including the QM atoms and MM residues within 10 Å of the QM region. The Hessian is mass-weighted and diagonalized to obtain vibrational frequencies. The frequencies change upon isotopic substitution, and from these changes, one calculates the KIE from Bigeleisen’s equation. We also evaluated potential of mean force (PMF) profiles of the transphosphorylation reaction, seen in Figure 1. The PMFs are performed from a series of simulations that differ only in their use of umbrella window restraining potentials to retain sampling at the desired location of the reaction coordinates. During these simulations, the values of the observed reaction coordinates are tabulated. Given knowledge of the umbrella window restraining potentials and the observed distribution of reaction coordinates, the underlying free energy surface can be calculated using the variational free energy perturbation method.

**RESULTS:**
We predict twister ribozyme has even less pronounced inverse O2′ and normal O5′ KIEs than in other ribozymes, such as RNase A. The KIEs corresponds to a transition state that has a slightly less fully formed O2′-P bond, and less fully broken P-O5′ bond with significant degree of proton transfer to the leaving group. This is consistent with the PMF profiles shown in Figure 1. The activation of the general base (its reception of the O2′ proton) is uncoupled from the transphosphorylation reaction coordinate; however, the motions of the general acid proton and transphosphorylation are concerted. Moreover,
the calculated intrinsic barrier of the transphosphorylation reaction (9.63 kcal/mol) is comparable to the intrinsic rate of much faster protein enzymes, such as RNAse A. These results help to answer and long-standing question in mechanistic enzymology. Molecules of RNA react more slowly because the catalytically competent state is much less probable than in protein enzymes; however, once this state is realized, the intrinsic rate of catalysis for RNA enzymes is comparable to that of protein enzymes.

**ROLE OF CALIBURN AND RDI:**

The Caliburn system has proven to be an ideal platform for developing and testing new software due to the community usage and administrative policies. Moreover, the *ab initio* QM/MM implementation is CPU-only, but is well parallelized, and it can take advantage of multiple nodes. Other available resources, such as Perceval and XSEDE computers, are used when appropriate, such as when GPU MD simulations are necessary.

**GRANTS AND PUBLICATIONS:**

This work was supported by the National Institutes of Health grants GM107485 and GM62248 in collaboration with Prof. Darrin M. York.


---

**Figure 1.** Left. Schematic of the transphosphorylation reaction in twister ribozyme. The general base proton moves from O2' of U-1 to the N-position of G33. The O2' of U-1 then attacks the P, breaking the bond with O5' within residue A1, and the protonated N-position of A1 donates the proton to the O5'. Right. Two dimensional PMFs. The first PMF illustrates the coupling between the movement of the general base proton (R1 – R2) and the transphosphorylation coordinate (R3 – R4). The second shows the coupling between the transphosphorylation coordinate and the movement of the general acid proton (R5 – R6).
Computational Investigation of Coke Formation on Platinum Alloy Catalysts

**PI:** Fuat Celik; **Student:** Jinwoong Nam, PhD Student  
**Institution:** Rutgers University–New Brunswick, Dept. of Chemical and Biochemical Engineering

**EXECUTIVE SUMMARY:**

Human activity often motivates the continued use of petrochemical fuels. That gas fuels have low cost for storage and comparatively high energy density make them useful for home heating and cooking. One of the challenges in the extraction process of gas fuels is in using separation processes that can extract specific components while separating others that are less desirable. The result is often an inefficiency that translates into economic cost, as well as environmental pollution, since the most economical solution to the trouble of flue gas is to flare them into the ambient air.

Catalytic dehydrogenation of these light alkanes can produce valuable olefins, commodity chemicals used in the synthesis of polymers. The materials that catalyze the dehydrogenation reaction are also responsible for significant coke formation, which leads to the build-up of carbon and deactivation of the catalyst, leading to costly down-time and cleanup operations. We use computational modeling to understand how the atomic structure of catalytic metal alloy particles leads to this undesirable side-reaction and how to prevent it.

**RESEARCH CHALLENGE:**

We will develop a fundamental understanding of the interaction between surface species and Pt-alloy dehydrogenation catalysts that give rise to their selectivity towards alkenes and for coke deposition during the high-temperature dehydrogenation of light alkanes. While the formation of the alkene has been studied, the mechanism of coke formation is unknown. We aim to generate potential energy surfaces and kinetic activation barriers for alkane dissociation reactions to carbonaceous surface species using Density Functional Theory (DFT) that will reveal how this mechanism is influenced by the electronic and geometric structure of the catalyst, and exploit these insights to design new catalyst alloy structures and compositions with enhanced resistance to deactivation.

Light alkanes, such as ethane and propane, produced as byproducts of hydrocarbon processing and found naturally as minority components of natural gas have little value, and are sometime flared, releasing CO₂ into the atmosphere. Dehydrogenation of these hydrocarbons to their corresponding alkenes would produce high value chemical and polymer precursors, and thus divert these greenhouse gas emissions.

**METHODS AND CODES:**

Density Functional Theory allows us to calculate the ground-state electronic structure of a many-electron system and makes the many-body Schrödinger equation tractable (which otherwise can only be solved for systems of few electrons due to severe computational resource requirements). Firstly, the Born-Oppenheimer approximation allows the nuclei (or nuclei plus core electrons) of all atoms to be treated with classical mechanics (rather than quantum mechanics), reducing their interactions with each other and with the valence electrons to electrostatic interactions. The primary basis for DFT is the Hohenberg-Kohn theorem, which states that the ground-state properties of the many-electron system are uniquely determined by the electron density, which is itself a function of spatial position. This defines an energy functional such that the correct ground state electron density minimizes the energy functional. Kohn-Sham DFT then reduces the intractable problem of many interacting electrons into a system of non-interacting single-electron wavefunctions that interact only with the effective potential arising due to all electrons (or rather the density of all electrons) in the system. Using DFT, we may calculate the electronic structure and total energy for the electronic ground state of a system with atomic nuclei in arbitrary positions, and algorithmically solve for the relative spatial positions of all atoms in the system that minimizes the total energy, i.e. a thermodynamically stable state.
We carry out periodic, self-consistent DFT calculations using the Vienna **Ab Initio** Simulation Package (VASP). The exchange-correlation functionals are calculated using the generalized gradient approximation (GGA-PW91) and projector-augmented wave (PAW) total-electron potentials. All calculations involving paramagnetic species are carried out with spin polarization. The single-electron wave functions are expanded using plane waves with an energy cutoff of 400 eV. A vacuum layer of 12 Å is used to separate the slabs in the $z$ direction, a dipole correction is applied, and the electrostatic potential adjusted to ensure the interactions between the surface of one slab and its periodic images are negligible. For a (2x2) surface unit cell, the Brillouin zone is sampled using a (6x6x1) Gamma-centered Monkhorst-Pack k-point mesh after checking for convergence with respect to increasing the number of k-points. Larger surface unit cells require fewer k-points (e.g. 4x4x1), and smaller unit cells require more k-points. This is because k-point sampling occurs in inverse or k-space, so larger distances in real space become closer together in inverse space, requiring fewer k-points.

**RESULTS:**

The effect of Sn depletion in the bulk of a PtSn alloy particle during ethane dehydrogenation was investigated by comparing the reaction pathways on Pt$_3$Sn/Pt(111) and Pt$_3$Sn(111). The binding energies of C$_1$H$_x$ and C$_2$H$_x$ radicals were generally stronger on Pt$_3$Sn(111) than on Pt$_3$Sn/Pt(111). Despite the electronic effect of neighboring Sn atoms, which typically weakens adsorbate binding energies, the expanded lattice constant in the bulk alloy leads to stronger binding energies.

**ROLE OF CALIBURN AND RDI²:**

Sampling all possible atomic configurations of systems of coadsorbates on metal surfaces requires considerable computational power. Calculating transition state structures on reaction pathways is even more demanding. The world-class computational power in RDI²’s high performance computing clusters provides ample resources for thorough computational chemistry investigations.

**PUBLICATIONS AND DATA SETS:**

Massively Parallel Optimization Methods in Machine Learning and Decision Analysis

**PI:** Jonathan Eckstein; **Student:** Gyorgy Matyasfalvi, PhD Candidate*
**Institution:** Rutgers Business School, Management Science & Information Systems

**EXECUTIVE SUMMARY:**

High Performance Computing (HPC) addresses some of the most challenging computational problems we face today. The key to harnessing the computational power of HPC systems, among others, lies in software’s ability to scale well. Existing work on implementing classical optimization methods on such computer architectures has been somewhat limited.

The project: Object-Parallel Optimization Software (OPOS), which has heavily relied on RDF resources, addresses this issue. The goal of OPOS is to provide the scientific community with reusable, customizable, and efficient implementations of classical optimization algorithms, that have well-known convergence properties, and scale well on HPC systems.

Power systems planning and machine learning are two fields that will benefit greatly from parallel optimization software and supercomputers, as many of the challenges in those fields boil down to solving large-scale optimization problems that are intractable with serial software and hardware. Consider the case of a power grid system where there are multiple users of energy (homes, businesses) that are spread out through a geographic region and there is a need to maintain a constant level of power on the grid system in order to ensure that enough power is provided, and no excess power is wasted. This problem requires having a good understanding of how the interplay between various sources and sinks for energy affect the overall load on a power system. Often, we do not know what the likely variance in power needs will be on a given day, adding a level of uncertainty to the problem that needs to be accounted for.

A popular problem in machine learning is classification, where one would like to assign labels to data points correctly. For example, classifying an email as spam or not-spam, or a news article as positive or negative in tone. One way of solving these problems is by solving a large-scale optimization problem to find a function that will best predict the label for unknown data points.

**RESEARCH CHALLENGE:**

Optimization problems can be difficult for several reasons: sometimes difficulty arises from the sheer size of a problem – for example, optimally allocating hundreds of resources to account for thousands of possible scenarios. In other cases, it may simply require a large amount of computation to verify that a particular solution is the best possible solution. The challenge lies in discovering algorithms that can fully utilize the computational power of supercomputers, thus speeding up computations so that large-scale optimization problems can be solved in a reasonable time.

**METHODS AND CODES:**

Software that can exploit the computational power of Software that can exploit the computational power of massively parallel supercomputers employ parallelization tools such as MPI, CUDA, or OpenMP. However, these tools require significant amounts of code “clutter” because they are linked to particular classes of hardware and are organized around relatively low-level operations. The same underlying algorithm may have to be re-implemented multiple times to adapt to different hardware environments or applications, and the resulting code may be difficult to read. A natural route to more elegant and portable implementation of parallel algorithms is to use established object-oriented programming concepts.

OPOS is implemented in C++ and makes heavy use of object-oriented programming. It includes many classes that allow for a clear, concise, reusable and efficient implementation of optimization methods. In particular, the algorithms are implemented via abstract classes. The result is clear, MATLAB-like yet efficient code that functions as an algorithmic template, readily applicable to various hardware platforms and data representations without the need for modification.

Within OPOS, one achieves concurrency by parallelizing optimization algorithms’ underlying linear algebra operations. In this context, it is essential to employ optimization algorithms that utilize only simple linear algebra operations that are relatively easy to implement in parallel.

**RESULTS:**

We tested OPOS on the continuous relaxation of a large-scale power systems planning problem. For these types of problems, the most expensive linear algebra operations that OPOS requires are matrix-vector multiplications, which are efficiently parallelized. The following figure illustrates a weak-scaling graph of the runtimes.

The graph was obtained by increasing the number of processors proportionally to the problem size (number of scenarios), beginning with 9 and ending with 576. The graph shows that OPOS exhibits good weak scaling properties. This is confirmed by our next table, extracted using Paraver® profiling software, which shows that OPOS spends most of the time (~90%) running floating point operations and little time (~10%) on communication.

For our text classification problem, we have developed a novel technique for matrix-vector and vector-transpose-matrix multiplication in distributed-memory parallel computing environments, where the matrices are unstructured, sparse, and have a substantially larger number of columns than rows or vice versa (which is

* Current position: Research Software Engineer, Princeton University
Our method allows for parallel I/O, does not require extensive preprocessing, and has the same communication complexity as matrix-vector multiplies with column or row partitioning. As the figure below shows our method achieves good strong-scaling results for matrices whose structure is disadvantageous for traditional column partitioning methods.

The figure above right is a strong scaling graph for two techniques (NzP stands for our novel nonzero partitioning technique, whereas ColP for the traditional column partitioning technique), with the number of processor cores appearing on the horizontal axis and the program running time, including I/O time, on the vertical axis. Clearly, the NzP algorithm exhibits much better scaling than ColP, since it does not have to contend with workload imbalances between processor cores. NzP exhibits good scaling behavior through 128 processor cores, after which little further speedup is obtained. It is important to note that at this point, the total computational time using 128 processor cores is less than 1 second for 1,000 multiplication pairs.

ROLE OF CALIBURN AND RDI:
There are two main reasons why we have chosen Caliburn for our computational runs. The first reason is the good relationship that already existed between Dr. Rodero and Gyorgy Matyasfalvi. Dr. Rodero’s mentoring allowed Gyorgy to quickly acquire the necessary parallel computing skills that were required for Gyorgy’s dissertation project. Secondly, getting fast access to a cutting-edge machine with fast interconnect and large number of cores/node was of paramount importance when it came to making good progress on our project. Caliburn allowed us to test the scalability and accuracy of our optimization code. Close collaboration on installation and optimization helped us to study possible bottlenecks using the profiling tools and get solutions at the fastest possible rates.

PUBLICATIONS AND DATA SETS:

External Software:
†Paraver (https://tools.bsc.es/paraver)
Catalytic Strategies of RNA Enzymes

**PI:** Darrin M. York; **Personnel:** Colin Gaines and Ken Kostenbader, Graduate Researchers; Solen Ekesan and Abir Ganguly, Postdoctoral Associates; Timothy J. Giese and Tai-sung Lee, Associate Research Professors; **Institution:** Rutgers University–New Brunswick, Department of Chemistry

**EXECUTIVE SUMMARY:**

The last four decades have witnessed considerable growth in our understanding of the functions of ribonucleic acid (RNA) in living organisms. During that time, the original notion that RNA molecules only serve as messenger intermediates in the pathway from the genetic code to protein synthesis, has been completely transformed. RNA is now known to be integral to a tremendous expanse of biological processes, such as regulation of gene expression and signaling pathways and catalyzing important biochemical reactions, including protein synthesis itself.

The remarkable ability of RNA molecules to catalyze complex chemical transformations has had profound impact in our understanding of the role of RNA in biology, the design of new biotechnology, and the formulation of theories into the origin of life. From a chemistry perspective, it is of fundamental interest to understand how RNA, with its limited repertoire of fairly inert functional groups, can achieve rate enhancements typically up to 6 orders of magnitude or more, in some cases comparable to the intrinsic rates of protein enzymes. A predictive understanding of catalytic RNA mechanisms may enable general principles to emerge that are transferable to the design of synthetic systems, such as the recently reported Hachimoji DNA/RNA, with great promise for new biotechnological applications.

Much of what is known about RNA catalysis is gleaned from nucleolytic ribozymes, a class of small catalytic RNA molecules that perform site-specific phosphodiester cleavage reactions and are involved in diverse, important biological processes. The study of nucleolytic ribozymes enable us to gain new insights into the mechanisms of RNA catalysis and facilitate the design of new RNA-based technologies.

Our research is application-driven, using state-of-the-art multiscale simulation methods to gain deep mechanistic insights into a series of nucleolytic RNA enzymes (ribozymes). We employ computational techniques which include large-scale atomistic molecular dynamics (MD) simulations to study the conformational landscape accessible to RNA molecules, GPU-accelerated alchemical free energy simulations to predict metal ion and cofactor binding and interpret experimental activity-pH profiles, and multidimensional ab initio QM/MM hybrid quantum mechanical/molecular mechanical (QM/MM) free energy simulations to elucidate catalytic pathways in enzymes and ribozymes, predict reaction rates, and explain the origin of chemical modification and mutational effects.

The results of these studies provide new insight into the function of RNA enzymes that can be used to aid in the design of new biotechnology and therapeutics.

**RESEARCH CHALLENGE:**

We investigated the catalytic mechanisms of a series of nucleolytic ribozymes, namely the twister ribozyme, the twister sister ribozyme, the Varkud satellite ribozyme, and the pistol ribozyme, and an artificially engineered DNA enzyme, the 8-17 DNAzyme. MD simulations were used to study the structure and dynamics of the ribozymes, GPU-accelerated free energy calculations were performed to calculate nucleobase $pK_a$s and simulate activity-pH profiles, and ab initio QM/MM simulations were performed to calculate multidimensional free energy surfaces and predict the catalytic reaction pathways. Our research provided new insights into ribozyme chemistry, including evidence for inactivity arising due to crystal packing in twister ribozyme, evidence of a novel role for divalent metal ions in Varkud satellite ribozyme, and indications that catalysis is supported by multiple metal binding modes in pistol ribozyme.

**METHODS AND CODES:**

All simulations made use of the AMBER18 (Lee et al, J Chem Inf Model, 58, (2018), 2043-2050) suite of programs that is partly developed at Rutgers, including the most recent GPU-accelerated free energy simulation capability (Lee et al, J Chem Inf Model, 58, (2018), 2043-2050, Lee et al; J Chem Theory Comput, 13, (2017), 3077-3084; Giese et al, J Chem Theory Comput, 14, (2018), 1564-1582), and has been widely tested on the local and national supercomputing facilities. The York Group are active AMBER developers and have built custom software for performing ab initio QM/MM simulations that make heavy usage of MPI libraries (MPICH2, Openmpi, etc.), and can run efficiently using the GNU Fortran, C, and C++ compilers and can further benefit in demanding cases from Intel or PGI compilers.
RESULTS:

Varkud satellite ribozyme (VSr)

VSr is the largest nucleolytic ribozyme and while it was one of the first catalytic RNAs to be discovered, it had not been previously studied using computational methods. We performed exhaustive MD simulations on different crystal structures to predict the active state of the ribozyme (Figure 1D). Our study provided evidence for a catalytic Mg$^{2+}$ ion that was previously unknown. In absence of the Mg$^{2+}$ ion simulations from all the crystal structures reflected an inactive catalytic site (Figure 1A), while in presence of the Mg$^{2+}$ ion simulations departing from the different crystal structures converged on an active site that is primed for catalysis (Figure 1B). We also performed multidimensional ab initio QM/MM free energy simulations to calculate the pathway and the free energy barrier associated with the catalytic reaction (Figure 1C). Our studies provided evidence of a novel role of divalent metal ions in ribozyme chemistry.

Figure 1. Computational investigation of catalysis in VSr. Panels (A) and (B) are 2D scatter plots of in-line attack angle ($\tau$) and O2'-P distance ($r$) constructed from MD simulations of the reactant state in absence and presence of the active site Mg$^{2+}$ ion, respectively. Panel (C) illustrates the free energy surface underlying the catalytic reaction obtained from ab initio QM/MM simulations. The black dotted line represents the favored reaction pathway of the catalytic reaction on the free energy surface. Panel (D) illustrates the active state ribozyme inferred from the computational studies.

Twister ribozyme (Twr)

The catalytically active state and chemical mechanism of the recently discovered Twr (Figure 2A) has been the focus of debate in the literature, owing to a flurry of both experimental and computational studies. We have performed the first end-to-end computational study of this ribozyme in order to provide an atomically detailed model of the catalytically active state in solution (Figure 2D). Starting with crystal simulations and MD in solution, we arrived at a model (Figure 2B) from which the chemical steps of the site-specific cleavage reaction could reasonably proceed. Ab initio QM/MM simulations were then carried out departing from this active state in order to determine the mechanistic pathways through multidimensional free energy landscapes for the reaction. We then characterized the rate controlling transition state and made prediction about kinetic isotope effects and line free energy relations. Ultimately, this work enabled us to discover that the structural scaffold of the twister active site (Figure 2C) conforms to the L-platform motif, a common catalytic core seen across a wide range of ribozyme classes including VSr, Pistol ribozyme and the 8-17 DNAzyme.

Figure 2. The structure of Twr and active site L-platform motif. (A) The sequence and secondary structure of Twr (PDB ID: 4OJI40), highlighting secondary structure elements (stems, loops, and pseudoknots). (B) Simulation snapshot showing the global fold of the twister ribozyme in a catalytically active state in solution, with color scheme matching that in Panel A. (C) Block cartoon showing active site base pairing that forms the L-platform motif for the twister ribozyme. The general acid and base are highlighted in red and blue respectively, with the scissile phosphate in magenta. The bolded residues (L1-I and residues 32-34) form the ‘L’ of this motif, while A2 and Y3 (gray) constitute the “L-anchor” that serves to anchor the general base. (D) Zoom in of snapshot from B highlighting the base pairing and hydrogen bonding around the scissile phosphate characteristic of the L-platform motif in twister active site. Residues depicted are the same as in C, with the addition of the phosphates of N16/17 shown anchoring A1 in the syn conformation.
Pistol ribozyme (Psr)

Psr is among the most recently discovered RNA enzymes, with exciting and unintuitive clues about its chemistry revealed by biochemical studies. To mesh this experimental data with the available crystallographic structures, which were at odds, we performed MD simulations on Psr to investigate the role of a solvated magnesium ion in the active site. With as much as four μs of sampling, we were able to convincingly demonstrate that Mg-bound water molecules not only supported catalysis (in the form of both charge transfer and proton transfer), but remained viable for catalysis throughout a broad range of at least six dissimilar binding modes (Figure 3). In addition to insights concerning the catalytic role of the active site metal ion, more insight about Psr’s striking similarity to the hammerhead ribozyme were uncovered through the use of long MD simulations.

Figure 3. Preferred catalytic strategies of Psr from MD simulations. The total heights of the stacked bars (left: 5K7C, right 5KTJ) depict the collection of MD frames that were observed with a catalytically relevant hydrogen bonding contact, classified as α, γ, β1, β2, or δ. Each bar is color-coded according to the percentage of this collection of catalytically relevant frames that contained a particular metal binding mode.

8-17 DNAzyme (8-17 dz)

The 8-17dz, a synthetic DNA molecule that has catalytic activity, is an archetype RNA-cleaving DNA enzyme. Departing from the recently published crystal structure (Figure 4), through molecular simulations and free energy calculations we have modeled the active state of the 8-17 DNAzyme (Figure 4 (inner panel)). Our results support a catalytic mechanism that bears striking similarity to features observed in the hammerhead and pistol ribozymes. The enzyme engineers an electrostatically strained active site that recruits monovalent and divalent metal ions to assist in active site organization and catalysis. These metal ions position the substrate for an in-line nucleophilic attack while facilitating the activation of the nucleophile by tuning of the general base pKₐ such that it can participate in catalysis and providing electrostatic stabilization to the leaving group.

Figure 4. Computational study of 8-17 dz. Crystal structure of 8-17 dz is illustrated with the enzyme strand highlighted in light gray, substrate strand in pink, and the catalytic site in blue. The inner panel illustrates the active state conformation of the catalytic site, with the general base shown in blue and the substrate cleavage site shown in pink. Pb²⁺ (gray) and Na⁺ (dark blue) ions help stabilize the active conformation.
ROLE OF CALIBURN AND RDI:

We run our MD simulations and free energy calculations on GPUs and QM/MM simulations on CPUs. ELF provides a valuable platform on which we can test our new GPU codes with short quick turn-around simulations. We typically run our production GPU simulations on XSEDE resources (BRIDGES GPU and COMET GPU) which have many more GPU nodes. Caliburn with its numerous CPU nodes and fast interconnect, provides a critical enabling technology for running our production QM/MM free energy simulations that involve running several (20-30) jobs requiring one node each.

PUBLICATIONS AND DATA SETS:


High Fidelity Simulations of Shock Wave Reflection Inside a Shock Tube

PI: Xianlian Alex Zhou, PhD
Institution: New Jersey Institute of Technology, Department of Biomedical Engineering

EXECUTIVE SUMMARY:
Exposure to blast waves is the leading cause of Traumatic Brain Injury (TBI) in military personnel. Many studies, both experimental and numerical ones, have been conducted to investigate the cause of TBI due to primary blast overpressure. Shock tubes are the main laboratory tools for generating blast-like shock waves traveling along a long chamber, within which different objects (such as animal or human phantoms and live animals such as rats) can be stationed and the overpressure reflected from and transmitted into the objects can be monitored. The Shock Wave Testing Facility at the Center of Injury Biomechanics, Materials, and Medicine (CIBM3) of New Jersey Institute of Technology (NJIT) has installed two square cross section shock tubes: one is 9” wide and the other is 28” wide, and both are longer than 6 meters. Work has been done to adjust and calibrate the shock tubes to generate pure, primary shock waves similar to that from field blasts. Currently, an ongoing effort has been carried out to investigate the reflection wave from objects placed inside the shock tubes. Plate objects with different sizes and material properties (e.g. aluminum vs plastic) have been tested and the pressure data on incident and reflection waves have been collected. To better understand and analyze the experimental data collected, we have been using modeling studies to reproduce some of the phenomenon observed and to reveal how sizes and materials affected the wave reflection. The findings will have profound impacts on how the shock wave shall be modeled correctly and accurately with data validation and how objects of different sizes and materials (rat vs human, phantom vs live tissue) could impact the wave reflection and transmission. And we expect this study will greatly help us to better understand TBI primary injury mechanism and continue to push the state-of-the-art in TBI research.

RESEARCH CHALLENGE:
There are several challenges in modeling shock wave traveling inside shock tubes with CFD (Computational Fluid Dynamics) and FSI (Fluid Structure Interaction) simulations. To capture the shock front with steep pressure gradient, very small cells must be used to obtain reasonable accuracy. Considering the length scale of the square shock tubes (~6m), the computational grid can easily exceed several million cells. In addition, a very small timestep, around the scale of 1μs (microsecond) or less, shall be used for transient/dynamic analyses. To simulate over 20ms of shock wave generation and propagation, typical workstation computers or even small clusters can take weeks or longer to complete just one simulation. Our goal is to conduct parametric simulations to cover most of the experimental conditions (different incident pressures and different objects). Consequently, high performance computing cluster resources are required.

METHODS AND CODES:
Simulations of shock wave reflection from an aluminum plate placed inside the 9” shock tube were conducted with an in-house multiphysics FSI solver that is written in C++ and uses MPI for parallel computation. The high-pressure chamber (at the right narrower end of Figure 1(a)) contains a mixture of helium and air whereas the square tube and the tapped transition region are filled with air at the atmospheric pressure. The square aluminum plate has a width of 3” and a thickness of 0.34” and is positioned in the middle of the square tube. Due to symmetry, only a quarter of the shock tube was modeled as shown in Figure 1(a). The mixture flow was solved with a finite volume based CFD solver and the deformation of the aluminum (linear elastic) plate was solved with a finite element solver. These two solvers were strongly coupled and iterated to achieve convergence during each time step.

RESULTS:
The FSI simulation of shock wave reflection solves several field variables, including the mass ratios of helium and air, flow pressure and temperature, as well as the deformation and stress of the plate. Many monitor points were placed inside the simulation domain to monitor the profiles of these variables. For example, the monitor point C1 is located on the top of the tube and in front of the plate; T4 is located behind the plate, as shown in Figure 1(b); and three monitor points are placed along the diagonal line of the plate. These monitor points are positioned at the same locations of pressure sensors employed in the experiments for model validation purpose. Figure 1(a-b) shows the pressure fields inside the shock tube at three different time steps right before and after the reflection, and Figure 1(c) shows the pressure profiles of the 5 monitor points for a 12ms duration. The reflected pressures on the plate are almost three times that of the incident pressure at C1. And the durations of positive pressure phases are between 4 to 5ms. These simulations provide the data needed for model calibration and cross-validation of experimental measurements and simulation methods and results. Additional simulations with plates of different materials and thicknesses can provide clues on how the pressure or impulse can be attenuated by the plates and if a plate of chosen material can be effective in reducing
pressure and protecting structures behind it. Through such a model validation study, we can gain confidence in applying this computational method to simulate the complex phenomenon of TBI with helmet protection. The predicted blast overpressure and positive pressure impulse (integral of pressure in time domain) are highly relevant indicators of potential TBI risks for people who sustained blast overpressure. Future work can be conducted to evaluate and design helmets with high efficacy in TBI protection.

Figure 1. The pressure of shock wave reflection from an aluminum plate inside the shock tube. (a) The full view of the shock tube; (b) zoom-in views of the shock wave reflection by the plate; (c) the pressure profiles at selected monitor points at C1, T4, and three locations on the plate.

ROLE OF CALIBURN AND RDI²:
The Caliburn cluster of RDF has cutting edge hardware and software for high performance computing. And the staff at RDF provided the expertise, help, and support we needed to conduct our simulations. It enables us to continue parametric simulations of shock wave reflection with different plate dimensions and materials and even totally different objects.
EXECUTIVE SUMMARY:

Supersonic and hypersonic vehicles have enabled a vast range of transport objectives. It was not that long ago that supersonic flights between London and Paris to New York were a regular occurrence. Our country’s ability to explore low earth orbit was also accelerated through our ability to navigate the upper atmosphere at high speeds in order to understand hypersonic flight conditions. This allowed for, save for a few tragic incidents, regular exploration and low gravity experimentation with the Space Shuttle.

A critical issue in the design of supersonic and hypersonic vehicles is the aerothermodynamic loading—the pressure, skin friction and heat transfer distribution over the vehicle surface. The pressure and skin friction determine the aerodynamic forces on the vehicle, and the heat transfer determines the thermal loads. Effective design of supersonic and hypersonic vehicles requires the development of accurate and efficient numerical simulation methods to predict the aerothermodynamic loading.

Our research focuses on the development of accurate simulation methods for modeling hypersonic flows over simplified vehicle geometries where shock waves interact with the flow near the vehicle surface. The accurate prediction of such flow interactions is critical to the effective design of hypersonic vehicles.

RESEARCH CHALLENGE:

The shape of a supersonic or hypersonic vehicle creates shock waves—extremely thin surfaces in the air across which the pressure, temperature and density increase. The interaction of such shock waves with the fluid layer—known as the boundary layer—near the surface of the vehicle can significantly affect the distribution of pressure and heat transfer on the vehicle. In particular, narrow regions of intense heat transfer can occur which may result in material failure and even vehicle loss. An example is the Flight 2-53-97 of the NASA X-15 hypersonic research aircraft on October 3, 1967. A dummy ramjet engine was attached to a pylon beneath the aircraft fuselage (Figure 1). During the flight at Mach 6.7 (i.e., 6.7 times the speed of sound) the shock waves generated by the dummy ramjet model and the main wing leading edge intersected to form a shock-shock interaction resulting in the formation of a high speed jet impinging on the dummy ramjet engine pylon and a significant increase in heat transfer. Within three minutes the high heat transfer completely disintegrated a portion of the pylon structure (Figure 2).

Current simulation methodologies—known as Reynolds-Averaged Navier-Stokes (RANS) methods—are incapable of predicting the aerothermodynamic loading due to hypersonic shock wave turbulent boundary layer interactions. A recent survey of methods demonstrated the inadequacy of RANS methods (Holden, M., Wadhams, T. and MacLean, M., Measurements in Regions of Shock Wave/Turbulent Boundary Layer Interaction on Double Cone and Hollow Cylinder/Flare Configurations from Mach 5 to 8 at Flight Velocities for Open and ‘Blind’ Code Evaluation and Validation, AIAA AVIATION 2014, June 2014, Atlanta, GA; Patel, S., Evaluation of k-omega Turbulence Model and Euler Flux Schemes for Shock Wave Turbulent Boundary Layer Interaction, MS Thesis, Department of Mechanical and Aerospace Engineering, Rutgers University, May 2018; Alviani, R., Assessment of Wilcox k-omega Turbulence Model in Regions of Shock-Wave Turbulent Boundary-Layer Interaction, MS Thesis, Department of Mechanical and Aerospace Engineering, Rutgers University, October 2018).

The research challenge is development of accurate simulation methods for hypersonic shock wave turbulent boundary layer interactions.
METHODS AND CODES:
Our approach is based on Large Eddy Simulation (LES). Unlike the RANS methods described above, LES incorporates the unsteady physics of hypersonic shock wave boundary layer interactions. We have developed a finite volume C++ code that solves the compressible nonequilibrium Navier-Stokes equations using the Implicit Large Eddy Simulation (ILES) method. Although ILES has been demonstrated to provide accurate simulation of shock wave turbulent boundary layer interactions at transonic and supersonic speeds, there is scant application of the ILES method to simulation of hypersonic shock wave turbulent boundary layer interactions including cold wall conditions. The anticipated outcome is a comprehensive assessment of ILES for simulation of hypersonic shock wave turbulent boundary layer (SWTBL).

If ILES methods are found to provide accurate prediction of hypersonic SWTBL, there will be an immediate and significant impact on the design methodology for hypersonic vehicles. In this sense, the term “accurate” means prediction of surface pressure and surface heat transfer within experimental uncertainty.

RESULTS:
Our simulations have demonstrated the capability for accurate modeling of supersonic turbulent boundary layer using ILES. Figure 3 displays an instantaneous static temperature iso-surface. The flow is from lower left to upper right. The iso-surface shows the turbulent structure of the boundary layer. Quantitative agreement is also achieved with known empirical correlations including mean velocity profile.

ROLE OF CALIBURN AND RDI: Caliburn provides a unique and critical resource for our simulations. The aforementioned supersonic turbulent boundary layer simulation required 384 core and 3.12 days wall clock time. Future simulations at higher Reynolds number and Mach number will require up to 1000 or more core and a week or more wall clock time. Caliburn is enabling us to explore new and exciting research in supersonic turbulent flow simulation.
Deep Learning of Protein-protein Interactions and Molecular Dynamics Simulations of the Orai Calcium Channel

**PI:** Guillaume Lamoureux  
**Institution:** Rutgers University–Camden, Department of Chemistry

**EXECUTIVE SUMMARY:**

At the microscopic level, many of the most important processes in biochemistry and biology are about molecular recognition. For instance, proteins in our body often acquire their biological activity by associating with one another and forming larger units like channels, transporters, motors, or sensors. Those nanomachines, in turn, may recognize other molecules as chemical signals, as sources of chemical energy, or as “cargo” to move across. What drives two molecules to bind and to work together? Can we predict which two molecules will “pick” one another among all possible interaction partners found in the crowded environment of a living cell? Our approach to the problem relies on high-performance computing and a detailed understanding of the physical forces that make two molecules attract or repel each other.

Our long-term goal is to be able to predict, through computation, whether any two molecules will bind or not and to reconstruct the network of molecular interactions present in any living organism. Not surprisingly, this network is perturbed by genetic mutations occurring in diseases like cancer and is key to understanding what makes an individual sick or healthy. Understanding the “rules” of molecular association will eventually allow us to use them to our own advantage, for instance to design antibody proteins with affinities for new antigens, or new drugs that target the proteins of pathogenic organisms.

This report describes two distinct projects for which our laboratory has received an RFD resource allocation. **Project #1** aims at using artificial intelligence techniques (namely, deep neural networks) to predict protein-protein interactions (PPIs). Two proteins are said to form a PPI if they bind to one another and assemble into a stable structure (either permanent or transient). PPIs are involved in numerous fundamental biological processes and a model that can reliably predict whether two proteins interact — and predict the effect of protein mutation/variation on an existing interaction — opens up new avenues for systems biology and for protein design. The project is done in collaboration with Dr. Yoshua Bengio, one of the leading experts in the field of deep learning.

**Project #2** aims at answering a fundamental question about the structural biology of the Orai subunit of the calcium-release-activated calcium (“CRAC”) channel, which is responsible for restoring and controlling the levels of calcium in the cell. Mutations of the CRAC/Orai channel are directly implicated in a number of conditions such as immunodeficiencies and muscle defects. Surprisingly, there is still no consensus about the stoichiometry of the Orai subunit. While the channel has been observed as a hexamer, with 6 identical copies of the Orai protein assembling into a pore, there are a number of equally compelling experimental evidences suggesting that the active form of the channel is a tetramer, formed of only 4 copies of the Orai protein. We have performed extensive molecular dynamics simulations of the channel in both its hexameric and tetrameric forms and have shown which structure is most consistent with experimental data.

**RESEARCH CHALLENGE:**

**Project #1:** Current state-of-the-art PPI prediction models rely on sequence similarity with proteins known to interact and have an intrinsically limited accuracy for the protein variants of interest to cancer biology and viral/bacterial infection. In simpler words, knowing that two proteins A and B interact does not guarantee that two similar (but different) proteins A’ and B’ will interact as well. The goal of the project is to train a deep learning model for the prediction of PPIs that relies only on the structural features of the individual proteins, and not on interaction data about similar proteins. The model we have designed predicts whether any two proteins A and B interact by learning the spatial features that, when convoluted, best predict the interaction and the structure of the AB complex. It uses three-dimensional (3D) convolutional networks, which are a generalization of the 2D convolutional networks used for image recognition. The model also uses a special network architecture that provides built-in rotational and translational invariance of the output, so that the A and B proteins can be presented in any orientation.

**Project #2:** While X-ray crystallography and cryogenic electron microscopy show that Orai proteins assemble as a hexamer, a number of functional experiments suggest that the Orai channel may function as a tetramer. Starting from the hexamer crystal structure, we had previously performed molecular dynamics (MD) simulations showing that a tetramer model of the channel was stable and had functional properties partially compatible with experiment. However, to resolve the longstanding “tetramer-versus-hexamer” debate, longer simulations of the “closed” and “open” structures are needed, so that we can observe how the stoichiometry of the channel affects its gating and its ion selectivity. Moreover, since the only structures available are for the “closed hexamer” conformation, considerable simulation time is needed to bring the “open hexamer”, “open tetramer”, and “closed tetramer” models to their relaxed conformations.
### METHODS AND CODES:

**Project #1** uses PyTorch and CUDA libraries, which are machine learning libraries optimized to build and train neural networks on GPUs. The project also uses our own TorchProteinLibrary, which is a computationally efficient library of differentiable primitives for deep neural network models of protein structure ([https://arxiv.org/abs/1812.01108](https://arxiv.org/abs/1812.01108)). TorchProteinLibrary implements the functionalities needed to perform end-to-end learning of protein structure prediction, including differentiable volumetric convolutions (based on the cuFFT library). The model also uses SE(3)-equivariant convolutional layers ([https://arxiv.org/abs/1807.02547](https://arxiv.org/abs/1807.02547)). **Project #2** uses NAMD, which is a widely used molecular dynamics code optimized for GPUs.

### RESULTS:

**Project 1:** The general architecture of the neural network is the following:

The spatial features $F_i$ are obtained from a rotationally and translationally invariant network. A number of architectures remain to be tested but, so far, the representation model that has shown the best prediction accuracy (higher than that of comparable rigid-docking methods such as ZDOCK and SwarmDock) is the following:

![Figure 1. Diagram of the overall model.](image)

**Project 2:** The extensive MD simulations and free energy calculations performed on the hexameric and tetrameric forms of the Orai channel have provided us with a much clearer picture of the activity of each conformation. To summarize, we find that the hexameric form of the Orai channel is more consistent with the calcium-selective function of the CRAC channel. We find that the Orai hexamer conducts Ca$^{2+}$ ions better than the Orai tetramer, but conducts K$^+$ and Na$^+$ ions more poorly. Since the CRAC channel is known to be selective against K$^+$ and Na$^+$ ions, our simulations provide a detailed description of the ion selectivity mechanism and support the hypothesis that the hexamer is the functional form of the channel.

### ROLE OF CALIBURN AND RDI$^2$:

The training of large neural networks (project #1) requires highly-available GPU nodes, which the RDI$^2$ ACI was consistently able to provide. MD simulations (project #2) run much faster on GPU nodes than on conventional CPU nodes, and the RDI$^2$ resources have allowed us to extend our simulations to the microsecond time scale (instead of the nanosecond scale typically accessible using CPUs).

### PUBLICATIONS AND DATA SETS:


Computational Study of Molecular Adsorption on MnO₂ nanoscaffolds

PI: Lu Wang; Postdoctoral Researcher: Gangotri Dey
Institution: Rutgers University–New Brunswick, Department of Chemistry

EXECUTIVE SUMMARY:
Certain kinds of injuries to the human body require extraordinary methods for partial or complete healing. Of particular concern are the regeneration of tissues like bone and neurons in areas of high-density structures, such as the lower limbs or spine, respectively. This is because these tissues do not have the natural capacity to divide and beget wholly new cells. Thus, we require novel approaches to generate these tissues.

An important avenue that modern medicine pursues in treating neurologically based injuries is through the injection of stem cells into injured sites, because stem cells have the capacity to very quickly beget the kinds of cells necessary to overcome the shortfalls in the natural characteristic of adult cells. For this purpose, researchers have designed nanoscaffolds that allow for the efficient delivery of drugs to facilitate the growth of the cells and keep their functions as they fill into tissues.

RESEARCH CHALLENGE:
Nanoscaffolds, which are multidimensional materials with nanometer length scale, have become promising platform for stem cell therapy and tissue engineering. In particular, these porous materials have remarkable surface areas and are excellent drug delivery agents to modulate stem cell differentiation and tissue regeneration. This work combines classical molecular dynamics (MD) simulations and density functional theory (DFT) calculations to uncover the adsorption mechanism of various neurogenic drugs on a novel biodegradable MnO₂ nanoscaffold, which has recently been designed and shown to be an excellent drug carrier to improve stem cell viability both in vitro and in vivo. To further elucidate the origin of their adsorption, we computationally investigate the binding of a series of small molecules with typical functional groups on the nanoscaffold. This study uncovers the key interactions involved in the adsorption process and provides guiding principles for the design of nanoscaffolds for efficient drug delivery in stem cell therapy.

METHODS AND CODES:
The first step of the study is to efficiently sample the configurations for drug binding on the MnO₂ nanoscaffold. We have used the software package GROMACS to perform MD simulations of drug adsorption using the universal force field. Each system contains a drug molecule, the MnO₂ nanoscaffold, water molecules and ions in a cubic box with a length of 10 nm. Each drug molecule comprises approximately 50 atoms and the MnO₂ surface contains 225 Mn and 450 O atoms.

After simulating the systems for 5 ns under constant temperature of 300 K and constant pressure of 1 bar, we have obtained the binding free energy surfaces and collected favorable adsorption geometries of the drug molecules on the MnO₂ surface.

From the configurations sampled from MD simulations, we have used the principles of DFT as implemented in the Quantum Espresso software package to study the electronic structures and adsorption energies of the drug molecules. For each system, we have performed geometry optimizations using the PBE functional along with the U parameters, which are included to treat the strong correlation in MnO₂ and dispersion corrections. Each optimization takes 8 nodes and approximately 120 CPU hours to complete. These calculations have also generated a few GB of scratch files that are later used to evaluate the density of states of the systems. The non-bonded interactions between the drug molecules and the MnO₂ surface are pictorially depicted using the NCIPLOT software.

RESULTS:
We have studied a set of neurogenic drug molecules: 1-azakenpaullone (AZP), rhodamine B (RhB), JQ1 and DAPT. By combining MD simulations and DFT calculations, we have identified that all drug molecules adsorb favorably to the MnO₂ nanoscaffold, although they have varying adsorption affinities (Figure 1). The optimized adsorption geometries of the drug molecules are shown in Figure 2, and the blue surfaces indicate that the adsorption sites are mainly at the aromatic moieties and polar functional groups.

The complex structures of the drug molecules make it difficult to disentangle the reasons for their different adsorption affinities. Hence, we have examined a series of small molecules with characteristic functional groups and revealed that molecular adsorption on MnO₂ nanoscaffolds is dominated by dispersion, electrostatic and charge transfer interactions. From this analysis, we have designed a metric that qualitatively predicts the adsorption affinity of a molecule and used it to elucidate the observed trends in the adsorption of neurogenic drugs (Figure 1). This metric is generally applicable to the adsorption of any guest molecules on the MnO₂ nanoscaffold, which will facilitate the experimental screening of proper adsorbates for efficient molecular delivery and aid the development of MnO₂-based nanoscaffolds for biomedical applications.
ROLE OF CALIBURN AND RDI2:

The parallel computing capacities of RDI2, in particular the Caliburn and ELF servers, have significantly facilitated our MD simulations of the MnO2-drug systems. As the DFT calculations are computationally highly demanding, they can’t be achieved without the computing resources and large memories of the RDI facilities.

PUBLICATIONS AND DATA SETS:

Direct Numerical Simulation of Blood Flow in Physiologically Realistic Microvascular Networks

**PI:** Prosenjit Bagchi  
**Institution:** Rutgers University–New Brunswick, Mechanical & Aerospace Engineering

### EXECUTIVE SUMMARY:

in human body refers to the flow of blood in vessels with diameters less than a few hundred microns. The majority of these blood vessels is known as the capillary vessels. They are responsible for gas and nutrient transport to surrounding tissues. They also play a major role in various diseases, such as, heart and brain diseases, kidney and retinal failures, and cancer. The vessels continually bifurcate into and merge with other vessels forming a geometrically complex 3D meshwork known as the microvascular networks. Distribution of blood in such networks dictates the pathophysiology of the surrounding tissue. Blood in microvessels flows as a concentrated suspension of discrete blood cells. The sizes of these cells are in the order of the size of the vessels. As such, the hydrodynamics of individual cell and their interaction with each other dictate the flow dynamics of blood in the microvascular networks. Furthermore, red blood cells (RBCs) are extremely deformable which allows them to squeeze through vessels with diameters much less than the cell size. There is no computational study of blood flow in the microvascular networks that faithfully accounts for the realistic and geometrically complex architecture of the networks while simultaneously and accurately resolving the deformation of every single RBC. This research has developed, for the first time, a high-fidelity, versatile, 3D model for accurately and efficiently modeling cellular-scale blood flow in the microvascular networks.

### RESEARCH CHALLENGE:

The problem of cellular-scale blood flow in complex geometry is a multiscale, multiphysics problem that involves diverse types of interfaces: the blood vessel walls represent the stationary rigid interfaces with highly complex shapes; the flowing RBC surfaces represent deforming interfaces; and the flowing white blood cell and platelet surfaces represent moving rigid boundaries. The problem also involves multiple fluids of different properties. The methodology must seamlessly integrate different interfaces, multiple fluids, and the physical laws that govern the evolution of these interfaces. RBCs which undergo extreme deformation and behave like soft matter exhibiting fluid/solid transition under varying shear rate of the imposed flow are much more complex than liquid drops. Accurate 3D modeling of individual RBC deformation, and extension of the model to dense suspension undoubtedly is a major challenge. Significant progress has been made in the past decade on modeling RBC deformation and their dynamics in dense suspension. To date, these studies have considered the flow in simple geometries. However, microvascular networks are geometrically very complex. As such, the RBC flow in the networks is expected to be very different from that in a simple geometry.

### METHODS AND CODES

The numerical approach utilizes Immersed Boundary Methods (IBMs) in the context of a 3D Finite Volume/Spectral fluid flow solver, and a Finite-Element solver for deforming interfaces. To accommodate the wide variety of interfaces, two separate types of IBMs are employed. The deformable cells are modeled with a continuous-forcing front tracking IBM. A Lagrangian mesh is fitted to the cell surface that moves with the flow and is used to calculate the forces in the membrane due to deformation using the Finite Element Method. These forces are then coupled to the bulk fluid via a body-force term added to the governing equations. For rigid interfaces, both stationary and moving, an explicit sharp-interface ghost node IBM (GNIBM) is developed. With this, a desired boundary condition at the interface is achieved by enforcing appropriate constraints on the fluid velocity at the Eulerian mesh nodes that are immediately outside the boundary (ghost nodes). For moving rigid interfaces, the GNIBM enforces constraints in accordance with the motion of the object.

### RESULTS

Figure 1 shows a snapshot from a representative simulation. As evident, the simulated networks are geometrically complex, and resemble the vascular architectures observed in vivo. While retaining this architectural complexity, at the same time, we resolve with high accuracy the deformation and dynamics of each individual RBC flowing through the network. Our simulations predict novel and unexpected phenomena that are absent in the simple geometry of straight tubes. One striking result obtained was negative pressure-flow correlations, implying a significant deviation from Poiseuille’s law. Negative correlations between hydrodynamic resistance and RBC volume fraction are also predicted, similarly defying a principle of particulate suspension.
ROLE OF CALIBURN AND RDI:

The simulations are computation- and data-intensive. The resolution used to discretize the flow domain is about 160 million points, plus 1000 cells each discretized using more than 5000 triangular elements. Each simulation generates 2-3 TB of data. Resources at RDI have been very useful for such computations.

PUBLICATIONS AND DATA SETS:

The study is published in Journal of Computational Physics, Journal of Fluid Mechanics, Physics of Fluids, Biophysical Journal, and Physiological Reports. It was funded by a NSF grant. Proposals are also currently pending with NIH and NSF.


Figure 1. A snapshot from a representative simulation of flow of deformable blood cells through a physiologically realistic network of capillary blood vessels.
Large Scale Simulations of Proteins and Nucleic Acids

**PI:** David Case  
**Institution:** Rutgers University–New Brunswick, Dept. of Chemistry and Chemical Biology

**EXECUTIVE SUMMARY:**

Nucleic acids play an important role as the building blueprints for all living things on Earth. These fascinating biomolecules contain within them the instructions needed to construct proteins that assist in the animation, as well as differentiation, of fauna and flora. While the small-scale structures of these nucleic acids are well known, their interactions as macromolecules are somewhat less understood.

To tackle the challenge of peering into the behavior of proteins and nucleic acids beyond their atomic composition, our group uses large numerical simulations to peer into the sub-structure of the larger scale molecules and test their behavior in the simulations with experimental observations.

**RESEARCH CHALLENGE:**

Our research focuses on the following areas of effort:

- **New approaches to computing chemical shifts in biomolecules.** We are developing and testing novel quantum chemistry models that break macromolecules into fragments and compute shielding tensors in a way that includes effects of the entire molecule and its solvent environment. Applications will include analysis of chemical shift patterns in helices and junctions of RNA and DNA, to conformational heterogeneity in proteins, to sequence-dependent shifts in collagen peptides, and to structure exploration in large viral RNAs.

- **Integral equation modeling of biomolecular solvation.** As an alternative to using molecular dynamics with explicit water molecules, the energetic and spatial behavior of solvent shells around biomolecules can be described by integral equation methods, which essentially implement classical (liquid) density functional theory. We previously incorporated this model into the Amber software suite, and have recently extended this approach (for the first time) to periodic systems, so that biomolecular crystals can be simulated. Our plan is to explore incorporating this model as an alternative to “bulk solvent” models in crystallographic refinement.

- **METHODS AND CODES:**
  
  Most MD simulations that our group run use a development version of Amber, see [ambermd.org](http://ambermd.org). Many simulations can be most efficiently carried out on GPUs, but large simulations, analysis, and (especially) integral equation methods still make use of (parallel) CPU facilities. A particular challenge for integral equation methods, which use a 3D-mesh surrounding the macromolecule, is the need for high precision (even extended precision BLAS codes) to obtain strict convergence over $10^6$ to $10^7$ grid points. We have integrated new versions of these codes into Amber, with recent extensions to periodic systems being of particular importance.

**RESULTS:**

Site binding of ions and water shapes nucleic acids folding, dynamics and biological function, complementing the more diffuse, non-specific “territorial” ion binding. Unlike territorial binding, prediction of site-specific binding to nucleic acids remains an unsolved challenge in computational biophysics. This work presents a new toolset based on the 3D-RISM molecular solvation theory and topological analysis that predicts cation and water site-binding to nucleic acids. 3D-RISM is shown to accurately capture alkali cations and water binding to the central channel, transversal loops and grooves of the Oxytricha nova’s telomeres G-quadruplex (Oxy-GQ), in agreement with high resolution crystallographic data. To improve the computed cations occupancy along the Oxy-GQ central channel, it was necessary to refine and validate new cation-oxygen parameters using structural and thermodynamic data available for crown ethers and ion channels. This single set of parameters that describes both localized and delocalized binding to various biological systems is used to gain insight into cation occupancy along the Oxy-GQ channel under various salt conditions. The paper concludes with prospects for extending the method to predict divalent metals binding to nucleic acids. This work advances the forefront of theoretical methods able to provide predictive insight into ion atmosphere effects on nucleic acids function.

The cyan “bubbles” in the center of the figure represent predicted positions for ion binding to a nucleic acid “quadruplex” in five discrete positions, which are in good agreement with experimental conclusions drawn from X-ray crystallography.

**ROLE OF CALIBURN AND RDI:**

The integral equation calculations discussed above use grids with $10^6$ to $10^8$ points, and require careful numerical analysis to achieve convergence, combined with large memories to hold multiple types of data on each grid point. Parallelization is limited by the need to repeatedly carry out three-dimensional fast Fourier transforms, but scaling to 16 threads is quite efficient, which makes these simulations well adapted to the existing Caliburn nodes.
PUBLICATIONS AND DATA SETS:

Using Parallel HPC to Determine Ramsey-Theoretic Quantities

**PI:** Doron Zeilberger, PhD  
**Institution:** Rutgers University–New Brunswick, Department of Mathematics

**EXECUTIVE SUMMARY:**

We conduct ongoing research in the area of Diophantine Ramsey theory, using computational methods to answer the question: for an equation \( E \) and an integer \( r \), what is the least integer \( N \) such that no partition of the integers 1 to \( N \) into \( r \) sets (often called “colors”) has a solution to \( E \) all in the same set (or same “color,” called “monochromatic”), regardless of how these \( r \) sets (“colors”) are assigned.

Diophantine Ramsey theory is based on Schur’s theorem, that any “coloring” of the integers 1 to \( N \), for \( N \) sufficiently large, there will be a monochromatic solution to the equation \( x + y = z \). The \( N \) required to guarantee this with two colors is \( N = 5 \). If we allow three colors, the required \( N \) is larger (it is 14).

HPC resources allow us to tackle the computation of Rado numbers simply by harnessing the computing power of more CPUs at once. While this is a slight oversimplification, it is essentially the parallelization of these algorithms that allows us to solve problems in hours or days that would otherwise take months or years, by doing exhaustive computations in parallel (best case, sometimes nearly perfect efficiency parallelization).

We use HPC methods to compute new Rado numbers for various equations widely considered to be of great interest but for which traditional non-HPC methods would have limited success. These results are, in and of themselves, meaningful progress, but they also provide insight into predicting and even proving bounds for larger families of equations and sometimes help reveal the underlying structure on which to base rigorous proof.

**RESEARCH CHALLENGE:**

Computing the size of extremal structures (the largest – or smallest, depending on the phrasing – structure of some type with – or without – a particular property) is often very difficult, and Ramsey Theory is no exception. There is a fact that the number of cases one might check grows at an extraordinary rate, making it very difficult to explore these structures concretely and exhaustively. For example, a graph on \( n \) vertices has \( O(n^3) \) edges, which might sound manageable, but \( O(r^{\omega(r)}) \) ways of splitting these edges up into \( r \) subsets and \( O(k^{\omega(r)}) \) potential subsets of size \( k \).

While a naive brute-force methodology is completely unfeasible, it is possible to exhaustively determine an upper bound for Rado numbers using other methods – taking advantage of new or existing methodologies for combinatorial search or similar problems, and implementing them in an HPC-aware way. Importantly, these methods must be effectively scalable to hundreds or thousands of cores/threads.

The back-of-the-envelope complexity of these systems can be described in several ways, depending on which methods are being used. For the remainder of this report, let \( r \) be the number of colors, \( E \) the equation in question (assume for simplicity it is polynomial), \( d \) the degree of \( E \) (assume for simplicity it is homogeneous in degree \( d \)), \( n \) the number of variables in \( E \), and \( N \) the Rado number (which is unknown).

There are roughly \( O(N^{(m-1)/2}) \) solutions to the equation to be checked, and a brute force approach requires checking every solution for each of \( N^r \) colorings. This superexponential growth must be put in check before we can advance further. We can, generally, describe other methods in comparable terms – number of solutions being checked in any particular coloring, and number of colorings being checked.

It is important to note that Rado numbers are not usually amenable to localized analysis because solutions to equations can span numbers from 1 to \( N \). Some special cases may be amenable to some decoupling of sub-intervals, particularly in the case of translation-invariant solution sets to very specific equations (e.g. arithmetic.
progressions). For this reason, we rely on parallelizable algorithms to solve these problems, and we consider two schemes – one which limits memory consumption more carefully, and the other that has greater potential to terminate in less time for the most difficult problems.

METHODS & CODE:

We have used high-performance computing, especially parallel-computing methods, to resolve this question and give this quantity for a number of different $E-r$ combinations. Our current research uses two types of algorithms, one a tree-search, the other a satisfiability-solver, to find these numbers. The cutting-edge computations depend on having as many processors as possible.

Using a depth-first search (DFS) of the tree of all colorings (also called “backtracking”) allows us to significantly reduce the computing time. We have written a custom package RADO, used to compute Rado numbers using such a highly-optimized tree-searching algorithm. This search is efficiently parallelized to properly manage memory usages, which is a strength of (non-parallel) DFS.

We will also make use of existing algorithms and libraries for solving problems of boolean satisfiability (“SAT-solvers”). We have written scripts that convert a Rado number problem into a complex set of Boolean clauses (in so-called conjunctive normal form). SAT-solving algorithms determine efficiently whether this is possible to satisfy these clauses, i.e. to avoid monochromatic solutions, providing the ability to compute Rado numbers.

The DFS search described in RADO, when parallelized, checks each branch of the search tree, and only has to check $O(n^{(m-1)/d})$ colorings, where $n$ is the length of the current branch. This alone represents a significant pruning of the search tree relative to brute-force.

The core of the RADO software package is a custom-built, parallelized depth-first tree-search. The algorithm scales one-to-one with the availability of nodes/cores using novel tree-pruning methodology and manages RAM effectively.

We have also implemented parallelized SAT-solvers (existing software, e.g. miniSAT) to compute Rado numbers that are not necessarily computable by our RADO program. SAT-solving resources are usually limited by RAM, more so than a DFS like that of RADO, but SAT-solvers have the potential to find Rado numbers that are a bit too hard for RADO, since in theory, SAT-solving is superior to DFS (in terms of time-complexity).

RESULTS:

We have successfully ported the RADO program to the HPC environments at RDF. We have adapted parts of RADO to bookend the work of a SAT-solver, allowing simple input like $E$ and $r$, generating the SAT clauses and handling the output automatically. This has allowed significant advances using both RADO and SAT-solvers, and we will continue to investigate hybrid approaches in the future.

One major goal of the project is to collect large volumes of data about some (relatively) easy to compute Rado numbers (from their relatively small certificates), but another is to compute Rado numbers for which bounds are not even known.

We will highlight two major findings that are made possible only by the generous grant of computing time at RDF. First, in the paper “Some Nonlinear Rado Numbers,” in the Electronic Journal of Combinatorial Number Theory (INTEGERS), we detail our computation of Rado numbers for equations of the form

$$x_1^2 + x_2^2 + \ldots + x_a^2 = y_1^2 + y_2^2 + \ldots + y_b^2,$$

which we might just call $E_{a,b}$. After compiling the following table of values, we were able to obtain a proof that proves a uniform bound for $a \leq b \leq c a$ for some constant $c$. Can you guess what the bound should be?

<table>
<thead>
<tr>
<th>$b$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>1</td>
<td>7825</td>
<td>105</td>
<td>23</td>
<td>18</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>23</td>
<td>17</td>
<td>26</td>
<td>17</td>
<td>34</td>
<td>14</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>19</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>15</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

The bound is proved to be 9 by a proof that is essentially illustrated by the following image, a proof constructed with the aid of computers as well, showing a search tree of all possible colorings. Each branch terminates with a set that can be used to form a monochromatic solution, assuming $a \leq b \leq c a$, where $c$ is dependent on the multiplicity of the numbers in that set needed to form that solution.
We have also proved a significant number of other results. In addition to many miscellaneous nonlinear Rado numbers, we have compiled a significant number of what are called generalized Schur numbers, which are for the following equation:

\[ x_1 + x_2 + ... + x_m + c = x_m \]

Using HPC resources (along with some lemmas to help us examine slightly simpler families of equations), we have been able to prove a significant number of 3- and 4-color generalized Schur numbers. We have formulated a closed-form formula which we conjecture to be correct except for cases with \( c \) and \( m \) small, and proved that this is at least a lower bound. (For comparison, beyond 4 colors, we know only a single Rado number, the 5-color Schur number, proved only in 2018 by similar methods using HPC by researchers at the University of Texas.)

**ROLE OF CALIBURN AND RDI²:**

Because our ability to solve these problems scales directly with the number of nodes/cores available and the time available, we are able to capitalize directly on the parallel processing capabilities of ACI resources to solve these problems in something like a \( \frac{1}{100} \)th or \( \frac{1}{1000} \)th fraction of the time on a conventional PC. We have benefited from the unique opportunity provided by RDI² to use its resources to accelerate our research.

**PUBLICATIONS, DATA SETS, & GRANTS:**

Caliburn is a compute system with state-of-the-art architecture that was built to significantly accelerate research computing performed by scientists and engineers in the State of New Jersey. For many first-time students and postdoctoral researchers, transitioning from working on serial computing to efficient high-performance computing can be a daunting task. In order to assist in the successful transition, members of the RDF staff were involved in three main service-directed activities.

1. Office hours – Regular office hours were established at the Rutgers–New Brunswick campus. Physical location was announced for a small conference room. Online links were provided and distributed widely to current users of Caliburn if they work in a different academic or research site.

2. Site visits were regularly scheduled for groups that were high-end users (frequently consuming over 90% of their awardee allocation near the end of their terms). This allowed for networking and greater interaction between members of the research team and the staff at RDF.

3. Visits with research staff that have particular research goals that require coordinated resource allocation. This might include consulting on queue submission strategies and designing workflows that take advantage of existing queue times, as well as use of existing drive space allocation. For certain projects, an evaluation of the fittingness of increasing queue times and drive allocation were also conducted.

ACHIEVING MAXIMUM BENEFIT FROM ADVANCED CYBERINFRASTRUCTURE. DEVELOPING A TWENTY-FIRST CENTURY STEM WORKFORCE.

Caliburn, the supercomputer at The Rutgers Discovery Informatics Institute (RDF), has enhanced the significance and scope of research and created new research and educational opportunities for Rutgers and New Jersey.

The Rutgers Discovery Informatics Institute operates an outreach and education program that reaches hundreds of individuals every year. These activities are aimed at members of the Rutgers community, New Jersey industry and academia, K-12 students and educators, and national researchers.

In addition to education and outreach events for advanced computing & research, RDF hosts the Distinguished Seminars featuring national and international experts, and an annual Open House event, a day of information sharing and learning through workshops and a poster showcase featuring the capacity and ability for Caliburn to enable research collaborations resulting in impactful discoveries not only for New Jersey but for people around the world.

The capacity and capability of the powerful supercomputer, Caliburn, has been promoted to the New Jersey community and beyond at multiple events and conferences, including the Caliburn launch event during Fall 2016, through the RDF website and the promotional video. Caliburn has also been featured in multiple articles and publications.
## RDI² Staff

### MANISH PARASHAR

Manish Parashar is Principal Investigator of the Caliburn project. He is Distinguished Professor of Computer Science at Rutgers, The State University of New Jersey University and the founding Director of the Rutgers Discovery Informatics Institute (RDI²). He is also Full Member (Clinical Investigations and Precision Therapeutics Program) of the Rutgers Cancer Institute, Visiting Professor in the Faculty of Business, Computing & Law at University of Derby, UK, and served as the Lead PI for Cyberinfrastructure for the NSF Ocean Observatories Initiative. At Rutgers, he co-led (with Prof. H. Berman) the strategic planning efforts in Research Computing and served as the Interim Associate Vice President of Research Computing between 2015–2016 to oversee the establishment of the Rutgers Office of Advanced Research Computing (OARC). He also co-founded the New Jersey Big Data Alliance (NJBDA), co-founded and was Co-Director of the Cloud and Autonomic Computing Center (CAC) NSF IUCRC at Rutgers (CAC@Rutgers) between 2008 and 2013.

Manish Parashar’s research interests are in the broad areas of Parallel and Distributed Computing and Computational and Data-Enabled Science and Engineering. Manish is the founding chair of the IEEE Technical Consortium on High Performance Computing (TCHPC), Editor-in-Chief of the IEEE Transactions on Parallel and Distributed Systems. He has received a number of awards for his research and leadership, and is Fellow of AAS, Fellow of IEEE/IEEE Computer Society and ACM Distinguished Scientist. For more information please visit [http://parashar.rutgers.edu/](http://parashar.rutgers.edu/).

Manish Parashar currently is on an IPA and currently serving as Office Director of the Office of Advanced Cyberinfrastructure (OAC) at the US National Science Foundation (NSF).

### PEGGY BRENNAN-TONETTA, PH.D

Peggy Brennan-Tonetta, Ph.D, is the Associate Vice President for Economic Development and responsible for the development and implementation of new university-wide economic development initiatives at Rutgers. In this role, she provides administrative leadership for the Rutgers Discovery Informatics Institute (RDI²). Peggy and Manish Parashar developed the State Higher Education Equipment Leasing Fund proposal which provided the funding for Caliburn, she served on the Procurement team to select the equipment providers for all phases of the project, serves as the liaison with the New Jersey Educational Facilities Authority for the expenditure and approval of funds for the project, and provided management support for the siting and installation of the modular data center and other aspects of the project. Peggy was a member of the team led by Prof. H. Berman and Prof. M. Parashar to develop a strategic plan for Research Computing at Rutgers and the establishment of the Rutgers Office of Advanced Research Computing (OARC). She is a co-founder and current President of the New Jersey Big Data Alliance. Peggy also led the development of the Rutgers’ Corporate Engagement Plan which resulted in the creation of the Office of Corporate Engagement and led the development of the Innovation Park® Rutgers Strategic Plan.

### IVAN RODERO

Ivan Rodero is Associate Director and Associate Research Professor at the Rutgers Discovery Informatics Institute (RDI²) at Rutgers, the State University of New Jersey, where he leads his research portfolio and provides leadership for the technical operations at RDI² and the establishment of the Advanced Cyberinfrastructure that includes Caliburn and impacts the strategy of the Institute. He has over 15 years of experience in cyberinfrastructure projects spanning from European-wide and US Federal funded projects, including leadership roles in constructing and operating the cyberinfrastructure for large science facilities such as the NSF Ocean Observatories Initiative (OOI) and research platforms such as the Virtual Data Collaboratory (VDC) in the US northeast region.

His research interests fall in the broad area of parallel and distributed computing and include high performance computing, energy efficiency, cloud computing and big data systems. His current research addresses new cyberinfrastructure models and aims at enabling the scalability and energy efficiency of next generation cyberinfrastructure including software-defined infrastructure co-design for big data analytics.

He has received various awards for his research and publications, including the IEEE TCSC Award for Excellence for Early Career Researchers (IEEE TCSC Young Achievers in Scalable Computing Award). He is senior member of IEEE, and member of ACM and AAAS.
J. J. VILLALOBOS
J. J. Villalobos is the Assistant Director of Research Computing and Cybersecurity at the Rutgers Discovery Informatics Institute (RDI), where he is responsible for the security stability and operational excellence of the research and production advanced cyberinfrastructure, which includes Caliburn, the Rutgers supercomputer, one of the fastest academic supercomputers in the United States, among the first clusters to use the Intel Omni-Path fabric and to equip its compute nodes with NVMe (non-volatile memory express) devices, making Caliburn a unique asset for the research community.

He joined RDI in January 2016, as Research Associate, and served as the technical lead for the National Science Foundation (NSF) Ocean Observatories Initiative cyberinfrastructure, a networked infrastructure of science-driven sensor systems that has transformed research of the oceans by integrating multiple scales of globally distributed marine observations into one observing system and allowing for that data to be freely downloaded over the internet in near-real time.

J. J. Villalobos is Co-PI for the NSF GeoSciFramework, a collaboration between computer scientists and geoscientists to develop a data framework for generalized real-time streaming analytics and machine learning for geoscience and hazards research. He also contributes as senior personnel in several federally-funded projects such as the NSF Virtual Data Collaboratory, a research platform in the US northeast region designed to drive data-intensive, interdisciplinary and collaborative research, and enable data-driven science and engineering discoveries.

J. J. Villalobos has over fifteen years of industry and research experience across several computer science and information technology disciplines including, but not limited to, high-performance computing, complex internet infrastructures and site reliability engineering operations.

PAUL ARIAS
Paul Arias is an Associate Research Scientist at the Rutgers Discovery Informatics Institute, where he is responsible for providing user support, technical consultation and policy advice for use of RDF systems.

Paul joined RDF in December of 2017. He was responsible for surveying the user community for Caliburn/ELF and wrote the user manual for the system, as well as coordinated regular meetings with top users to help tackle bottlenecks in the research output, lessons that were incorporated into new instructions for the system user manual. He also visited various campuses in NJ to promote accessibility of Caliburn to the wider research community and held online workshops to communicate Caliburn’s potential to accelerate research in the state of New Jersey.

FOROUGH GAHGRAMANI
Forough Ghahramani is Associate Director at the Rutgers Discovery Informatics Institute (RDF). In this role, she leads the educational programing focused on big data and large scale computing, and is responsible for vision and strategic planning for outreach and enhancing research productivity for the institute through partnerships, corporate engagement and cross-campus collaborations. In her marketing leadership role, Forough develops the marketing collateral for RDF and promotes Caliburn through educational programs, conferences, events, social media, news articles, and Research Computing Information Sessions at Rutgers and in the New Jersey Community. Forough has diversified experience in industry, higher education, and entrepreneurship. Her experience in higher education includes previously serving as faculty, dean, and department chair. Prior to joining academia, she held senior level software engineering and management positions at Hewlett Packard.

Forough has a doctorate in Higher Education Management from University of Pennsylvania, an MBA from DePaul University, MS in Computer Science from Villanova University, and BS in Mathematics with a minor in Biology from Pennsylvania State University. Her research interests include, bioinformatics, the opportunities associated with the convergence of biotechnology and information technology, Models for Fostering diversity and inclusion in the innovation life cycle.

Forough has extensive experience evaluating the policies and programs that shape vibrant innovation ecosystems, and is consulted on the state, national, and international levels in various capacities including STEM workforce development strategies. Forough serves in leadership roles in educational and non-profit boards.
Key Words/Topics

A
AMBER 14, 34
AMBER18 20
SANDER 14
Astrophysics 12

B
Boolean satisfiability (SAT-solvers) 37
Brain derived neurotrophic factor 10
M66 11
Met-Met 11
Val66Met 10

C
Calcium channel 28
Orai subunit 28
hexamer 28
tetramer 28
Catalysis 21, 22
dehydrogenation 16
for RNA 15, 20
Computation Fluid Dynamics (CFD) 24
Fluid Structure Interaction 24
Large Eddy Simulation (LES) 26
Implicit 27
Reynolds-Averaged Navier-Stokes (RANS) 26
shock wave 24
hypersonic 26
shock wave reflection 24
turbulent boundary layer 26

D
Density Functional Theory (DFT) 16
Hohenberg-Kohn theorem 16
Vienna Ab Initio Simulation Package 17
Depth-first search (DFS) 37
Diophantine Ramsey theory 36
Diseases
cancer 28, 32
kidney 32
Neurodegenerative 10
Retinal failures 32
DNA 20, 34
DNAzyme 20
synthetic 22

F
Fuels 16
petrochemical 16

G
Galaxy 12
evolution 12
Milky Way 12
GROMACS 10, 30

I
Immersed Boundary Method (IBM) 32

K
Kinetic isotope effect values (KIEs) 14

M
Machine learning 18
Microcirculation (blood) 32
Microvascular networks 32

N
NAMD 29
Nanomachines 28
Nanoscaffolds 30
MnO2 30

O
Object-Parallel Optimization Software (OPOS) 18
Optimization 18
classification 18
machine learning 18
power grid system 18

P
Parallelization libraries
CUDA 18
Message Passing Interface (MPI) 18
OpenMP 18
Paraver 18
Parallelization libraries
OpenMP 18
Platinum Tin Alloys (PtSn, Pt3Sn) 16
potential of mean force (PMF) 14
Power grid 18
Protein -protein interactions 28
structures 28

Q
Quantum Espresso 30
Quantum mechanics 16
molecular dynamics 20, 28, 30, 34

R
Rado numbers 36
Ramsey-theoretic quantities 36
RNA 14, 20, 34
phosphodiester cleavage reactions 20
Pistol ribozyme (Psr) 22
RNase 14
Twister ribozyme (Twr) 21
Varkud satellite ribozyme (VSr) 21
viral 34

S
Schur’s theorem 36
Spectral energy distribution 12
Star formation history 12
Stem cells 30

T
Telescopes
Chandra Satellite 12
Hubble Space 12
Large Synoptic Survey 12
Spitzer Space 12
Telomeres 34
TorchProteinLibrary 29
Traumatic Brain Injury (TBI) 24
TreeCorr 12
This publication was designed and produced by the Rutgers Discovery Informatics Institute (RDF), a division of the Office of Research and Economic Development at Rutgers University. RDF extends its thanks to the research teams who contributed summaries of their latest Caliburn enabled research for inclusion in this report. Caliburn is supported by the State of New Jersey and Rutgers University.