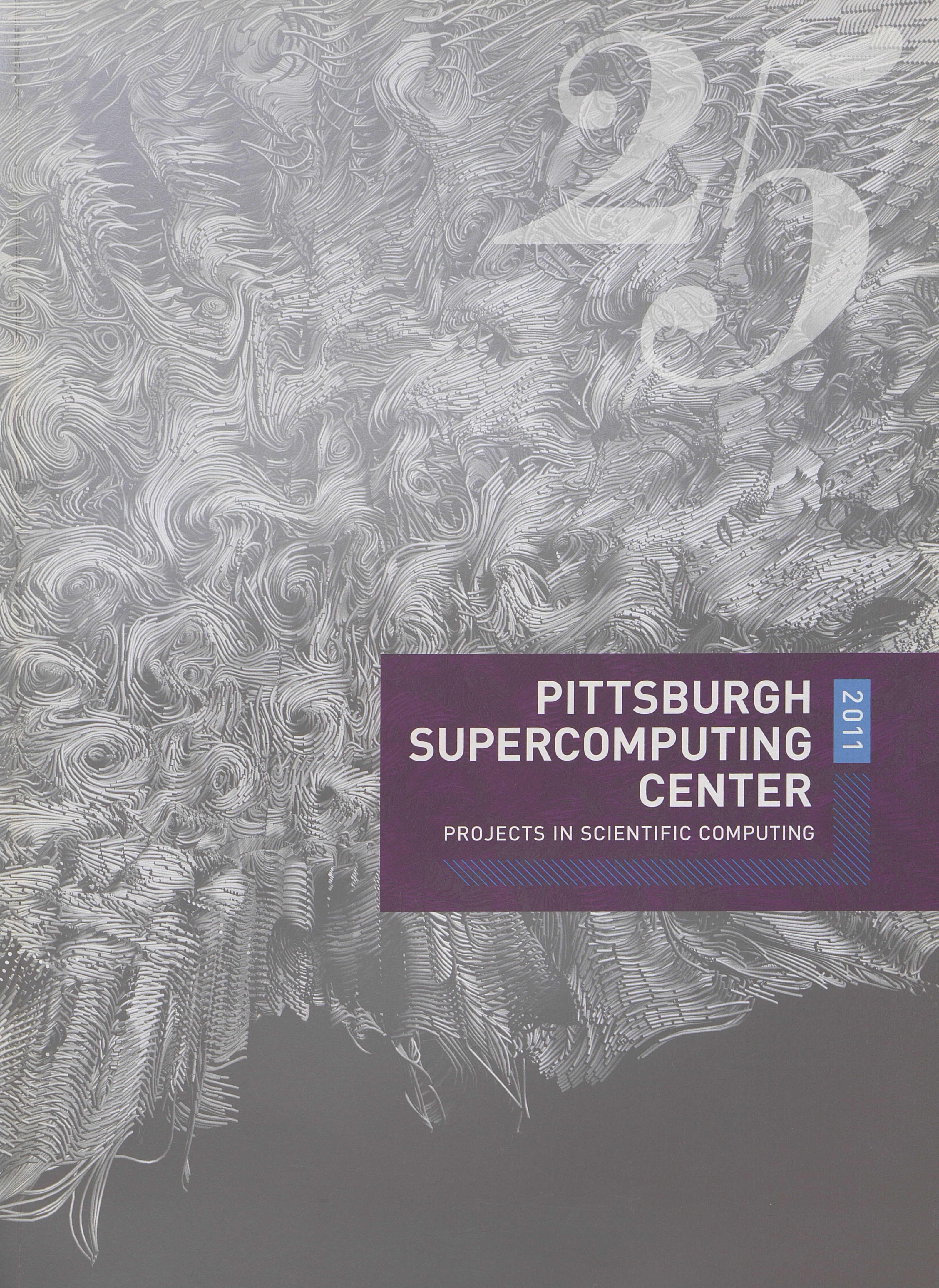
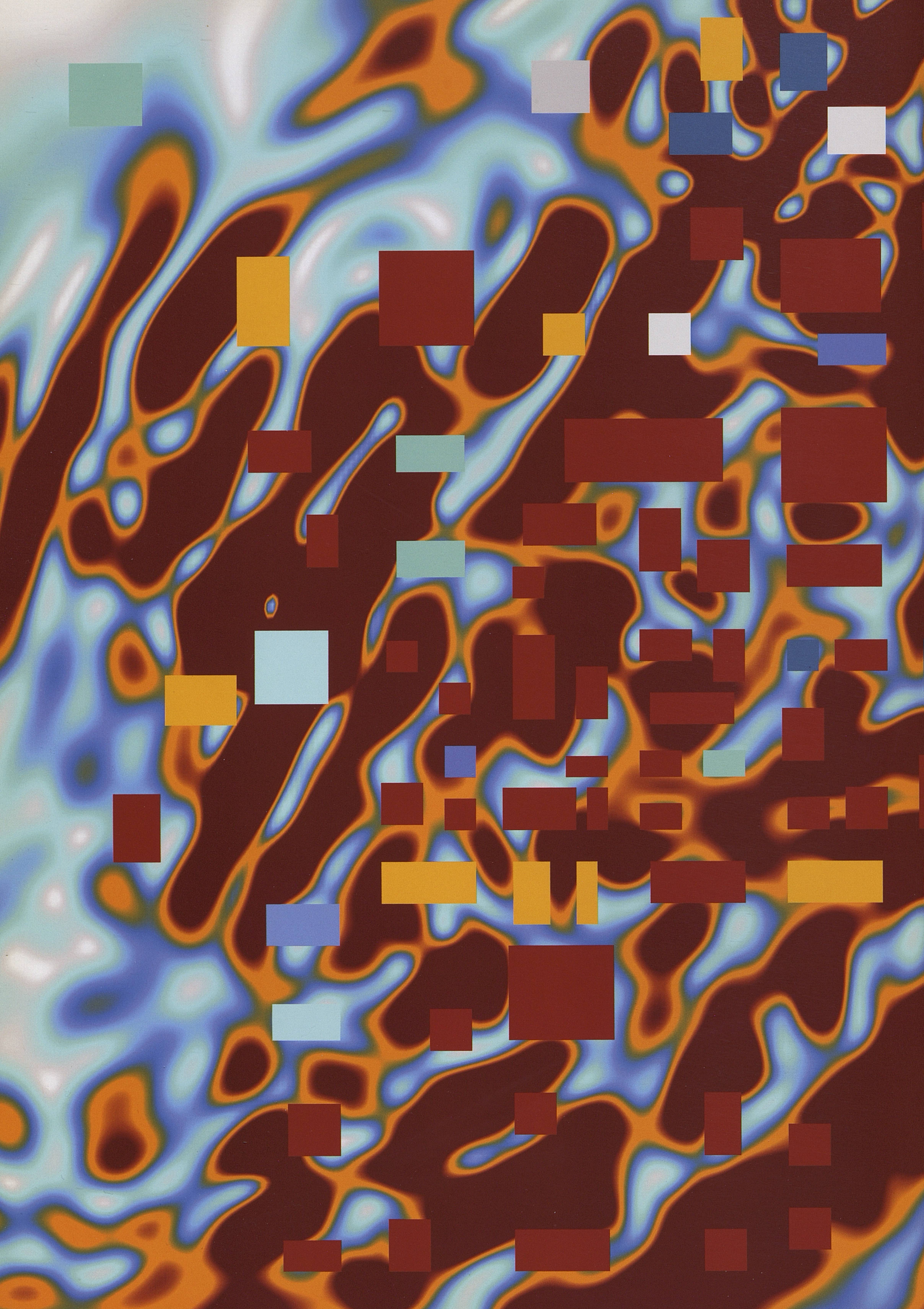


PITTSBURGH SUPERCOMPUTING CENTER

2011

PROJECTS IN SCIENTIFIC COMPUTING



PSC.EDU / 11

The Pittsburgh Supercomputing Center provides university, government and industrial researchers with access to several of the most powerful systems for high-performance computing, communications and data-handling available to scientists and engineers nationwide for unclassified research. PSC advances the state-of-the-art in high-performance computing, communications and informatics and offers a flexible environment for solving the largest and most challenging problems in computational science. As a leading partner in XSEDE, the Extreme Science and Engineering Discovery Environment, the National Science Foundation's cyberinfrastructure program, PSC works with other XSEDE participants to harness the full range of information technologies to enable discovery in U.S. science and engineering.

www.psc.edu | **412-268-4960**

FOREWORD FROM THE DIRECTORS

Twenty-five years ago, with a grant from the National Science Foundation, the Pittsburgh Supercomputing Center (PSC) was born. Since then we've participated in breathtaking technological change, and it's still happening. This year, with partner sites around the country, we embark on the adventure of XSEDE, the Extreme Science and Discovery Environment (p. 5), the most powerful collection of integrated computational resources in the world.

Two supercomputing systems that we host and make available to the national research community this year came into productive maturity. With Anton, the world's most effective system for simulation of proteins and nucleic acids, computational biologists have opened a new view into protein dynamics. We briefly describe in this booklet the unique story of Anton and findings from four of these projects (pp. 18-23).

Blacklight, the world's largest shared-memory system (p. 4), has rapidly become a force across a wide and interesting spectrum of fields — including genomics, machine learning, natural language processing, geophysics and astrophysics.

As a tool for assembly of sequence data from next-generation sequencing instruments, Blacklight enabled remarkably fast results in two projects (pp. 28-31). One, led by James Vincent, is the sequencing of an NIH model organism, a fish called the little skate. Similarly, Blacklight accelerated sequence assembly in the work of Cecilia Lo at the University of Pittsburgh Medical School on congenital heart defects.

With limitless quantities of text available on World Wide Web, Blacklight's shared memory is a powerful tool for sifting words as data — as Noah Smith showed in four papers in diverse fields of "natural language processing" (pp. 32-35) within six months of access to Blacklight.

For astrophysicists Tiziana Di Matteo and Rupert Croft, Blacklight has revolutionized discovery from large-scale simulations of how the cosmos evolves (pp. 40-43). The ability to hold an entire snapshot of MassiveBlack, their huge simulation, in memory at one time was instrumental in their ability to reveal "cold gas flows" as a phenomenon that accounts for supermassive black holes in the early universe.

With help from PSC scientist Marcela Madrid, Catalina Achim solved the structure of a fascinating molecule called peptide nucleic acid (pp. 36-39), a close cousin structurally to DNA, but with important advantages for research in electron transport.

In a major accomplishment, Art Wetzel and Greg Hood, scientists in PSC's National Resource for Biomedical Supercomputing (pp. 13-14), co-authored a paper that appeared as a cover story in *Nature*, the prestigious international journal of science. Their collaboration with Clay Reid and colleagues at Harvard (pp. 24-27) is a milestone in brain research.

Along with these scientific advances, PSC continues to be a resource for research and education in Pennsylvania (p. 6). Through the Three Rivers Optical Exchange (pp. 11-12), PSC's networking group serves the Pennsylvania-West Virginia region and carries out nationally recognized research in next-generation, high-bandwidth Internet resources. This booklet also highlights (pp. 9-10) our important work to help educate the upcoming generation of scientists and science-literate citizens.

Much more than technology *per se*, it's PSC's staff who make all of this possible. It's our privilege to work with a collection of people second-to-none in world-class talent and experience in high-performance computing. We're grateful also for support from the National Science Foundation, the U.S. Department of Energy, the National Institutes of Health, the Commonwealth of Pennsylvania and many others.

Ralph Roskies

Michael Levine (left) and **Ralph Roskies**, PSC co-scientific directors

CONTENTS

BLACKLIGHT GOES TO WORK

With a $2.8 million award from the National Science Foundation, PSC introduced the world's largest shared-memory supercomputer

Researchers are making productive use of Blacklight. This new system, which PSC acquired in July 2010, with help from a $2.8 million award from the National Science Foundation, has opened new capability for U.S. scientists and engineers. With 512 eight-core Intel Xeon 7500 (Nehalem) processors (4,096 cores) and 32 terabytes of memory, Blacklight is partitioned into two connected 16-terabyte coherent shared-memory systems — the two largest shared-memory systems in the world.

In computer terms, "shared memory" means a system's memory can be directly accessed from all of its processors, as opposed to distributed memory (in which each processor's memory is directly accessed only by that processor). Because all processors share a single view of data, a shared memory system is, relatively speaking, easy to program and use.

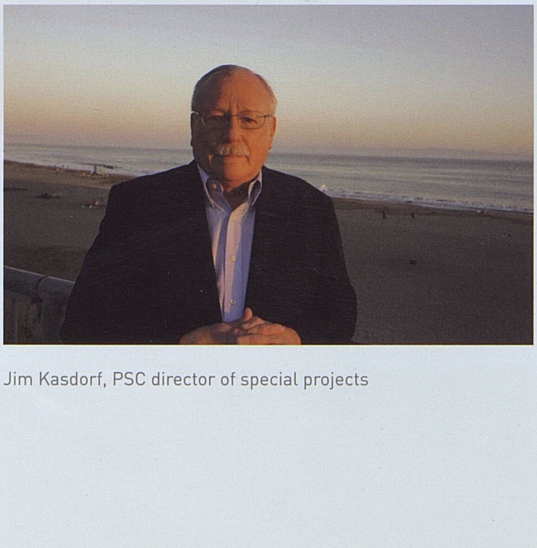
"For many research communities — including data analysis and many areas of computer science," said PSC scientific directors Michael Levine and Ralph Roskies in October 2010, as Blacklight became a production resource, "Blacklight opens the door to high-performance computation and thereby expands the abilities of scientists to ask and answer questions."

As described in this publication, Blacklight has already enabled advances in nanomaterials (p. 14), genomics (p. 28), machine learning (p. 32), astrophysics (p. 40), geophysics (p. 44), natural language processing, (p. 46) and climate modeling (p. 47).

Blacklight: The SGI® Altix® UV1000 system

BLACKLIGHT MEMORY ADVANTAGE PROGRAM

To help researchers take advantage of Blacklight, PSC provides a Memory Advantage Program to develop applications that can effectively use Blacklight's shared-memory capabilities. These include rapid expression of algorithms — such as graph-theoretical software, for which distributed memory often presents obstacles, and interactive analysis of large data sets, which often can be loaded in their entirety into Blacklight's shared memory. For such projects, a PSC consultant can provide advice on debugging and performance-analysis tools and procedures, and other fixes and optimizations. Interested researchers may contact: *remarks@psc.edu*.



Jim Kasdorf, PSC director of special projects

CREATING NATIONAL CYBERINFRASTRUCTURE

As a leading partner in XSEDE, the most powerful collection of integrated digital resources and services in the world, PSC helps to shape the vision and progress of U.S. science and engineering

PSC & XSEDE

Through XSEDE, the Extreme Science and Engineering Discovery Environment, the NSF cyberinfrastructure program that launched this year, PSC extends its active role in the development of national cyberinfrastructure. XSEDE replaces and expands on the TeraGrid, the predecessor NSF program that began more than a decade ago. More than 10,000 scientists used TeraGrid to complete thousands of research projects. Similar work — only in more detail and in a broader range of fields — continues with XSEDE.

PSC-scientific co-director, Ralph Roskies is a co-principal investigator of XSEDE and co-leads its Extended Collaborative Support Services (ECSS). "ECSS staff work both with user groups in fields familiar with high-performance computing," says Roskies "and with the XSEDE outreach team to reach user groups, communities and digital services that are new to HPC."

Other PSC staff lead many areas of the comprehensive XSEDE program. Janet Brown, who manages PSC's network research, leads the XSEDE Systems and Software Engineering team that oversees the software environment that integrates resources among many providers. As manager of XSEDE Outreach Services, PSC manager of education, outreach and training Laura McGinnis leads programs that help to prepare the next generation of computational scientists.

PSC's security officer, Jim Marsteller, is the Incident Response Lead for XSEDE. Wendy Huntoon, PSC director of networking, is XSEDE Networking Lead. Ken Hackworth, PSC's user relations coordinator, leads the XSEDE allocations process by which research proposals are reviewed and evaluated to receive grants of computational time on XSEDE resources. PSC scientist Sergiu Sanielevici, director of scientific applications and user support for PSC, leads the Novel and Innovative Projects area of XSEDE's ECSS effort, which focuses on development of projects in fields or from institutions and communities that can exploit advanced computing but haven't traditionally used it.

XSEDE
Extreme Science and Engineering Discovery Environment

XSEDE PARTNERS

University of Illinois at Urbana-Champaign

Carnegie Mellon University & the University of Pittsburgh

University of Texas at Austin

University of Tennessee, Knoxville

University of Virginia

Shodor Education Foundation

Southeastern Universities Research Association

University of Chicago

University of California San Diego

Indiana University

Jülich Supercomputing Centre

Purdue University

Cornell University

Ohio State University

University of California, Berkeley

Rice University

The National Center for Atmospheric Research

MORE INFO:
xsede.org

PSC's directors (l to r), who oversee day-to-day PSC operations and help to coordinate PSC's role in XSEDE: Cheryl Begandy, director, education, outreach & training; Bob Stock, PSC associate director; David Kapcin, director of financial affairs; Sergiu Sanielevici director, scientific applications & user support; David Moses, executive director; Wendy Huntoon, director of networking.

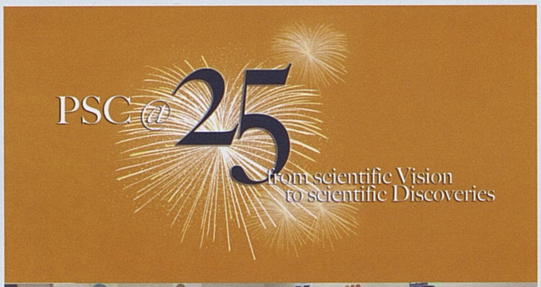
Not pictured: Nick Nystrom, director, strategic applications; J.Ray Scott, director, systems & operations.

SUPERCOMPUTING IN PENNSYLVANIA

With Commonwealth of Pennsylvania support, PSC provides education, consulting, advanced network access and computational resources to scientists and engineers, teachers and students across the state

DISCOVER 11: 25 YEARS OF PSC SERVICE

Over 100 students, representatives of government and industry participated in PSC's 25th anniversary observance and Discover 11 Open House on April 15. The event featured demonstrations of PSC research, including 3D stereo movies. PSC's biomedical group highlighted work published as a cover article in the March 10 issue of *Nature*, the prestigious international science journal (see pp. 24-27).



ECONOMIC IMPACT: RETURN ON INVESTMENT

A report by the economic analysis firm Fourth Economy this year detailed the economic impact that PSC has provided to Pennsylvania. Over the 25-year span since PSC's founding in 1986, $13.20 in federal funds have come to Pennsylvania for every dollar of the state's investment in PSC. In addition to attracting federal investment, PSC is a job anchor, providing more than 1600 jobs (direct and indirect) annually. PSC has also trained more than 300 employees now working at Pennsylvania companies.

Other benefits of PSC's presence include its talent pool as a magnet for other technology investments. PSC staff work was instrumental in the proposal that won $100 million in federal support for PennREN, a statewide network that, along with providing broadband connectivity to rural residents, businesses and agencies, will generate many jobs.

PSC PROVIDES TO PENNSYLVANIA:

- economic impact of $219 million annually,
- 1,666 jobs, including PSC employees and research partners who contribute $2 million annually in state payroll taxes,
- $450 million in federal investment, leveraged from $34 million in Pennsylvania investment, a return of $13.20 per dollar, and
- a vital resource for companies seeking to innovate.

PSC COMMONWEALTH ADVISORY COMMITTEE

A Commonwealth Advisory Committee helps PSC maximize its benefits to Pennsylvania. These community leaders help to integrate PSC's work with economic development, recommend new program areas, and promote PSC visibility:

The Honorable Jay Costa
Senate of Pennsylvania

The Honorable Michael Folmer
Senate of Pennsylvania

The Honorable Joseph Markosek
Pennsylvania House of Representatives

Thomas D Moser
Manager, Infrastructure and Network
Westinghouse Information Technology

David Shapira
Chief Executive Officer
Giant Eagle

Colton Weber
Technology Development Consultant
Pennsylvania Department of Community and Economic Development

Robert Wonderling
President and CEO,
Greater Philadelphia Chamber of Commerce

Dennis Yablonsky
Chief Executive Officer
Allegheny Conference on Community Development and Affiliates

RESEARCH & TRAINING AT PENNSYLVANIA COMPANIES, COLLEGES & UNIVERSITIES, 2009-2010

From July 2010 through June 2011, PSC workshops and presentations in high-performance computing reached 586 Pennsylvania grad and undergrad students, and PSC provided more than 2.3 million processor hours to 685 individual Pennsylvania researchers from 22 institutions. The following Pennsylvania corporations, universities and colleges used PSC resources during this period:

Bryn Mawr College
Carnegie Mellon University
Cedar Crest College
Cheyney University of Pennsylvania
Dickinson College
Drexel University
Duquesne University
Indiana University of PA, all campuses
Kutztown University of Pennsylvania
Lehigh University
Lock Haven University
Pennsylvania State University, all campuses
Shippensburg University of Pennsylvania
Swarthmore College
Temple University
University of Pennsylvania
University of Pittsburgh, all campuses
Ursinus College
Westinghouse Electric
Westinghouse Management Services, Inc.
Widener University
Wilkes University

3ROX: NETWORK FOR EDUCATION

The Three Rivers Optical Exchange (3ROX) provides research and education network service to six Intermediate Units in western Pennsylvania that serve 128 school districts, more than 800 schools, 25,000 teachers and 300,000 students. 3ROX links these schools, teachers and students to a global community of people and ideas.

PENNSYLVANIA RESEARCH INNOVATION

A number of projects in this booklet exemplify research by scientists in Pennsylvania:

Nanomappers of the Mind: PSC scientists help to pioneer a research approach that maps a "wiring diagram" of the brain (p. 24).

Putting Genes Together Really Fast: A University of Pittsburgh Medical School scientist leads a genomics study of congenital heart disease (p. 28).

Mining the Word Hoard: PSC's newest system, Blacklight, enables new studies in natural language processing (p. 32).

No Charge Double Helix: A team of scientists derive the first accurate structure of a fascinating molecule with applications in biomedicine and nanotechnology (p. 36).

Supermassive Growth Spurt: Astrophysicists solve a puzzle about the origins of the first black holes in the universe (p. 40).

Shutting the Door on HIV: A scientist at Bryn Mawr College explores a promising avenue for drug therapy to defeat AIDS (p. 45).

Jukebox with a Brain: Carnegie Mellon's Department of Machine Learning, the first such department in the world, used PSC resources in a prestigious competition. (p. 46).

ENERGIZING SCIENCE LEARNING

PSC programs in science education give the Pittsburgh region a jumpstart toward a cyber-savvy workforce

"Introducing 'cool' technology into the classroom engages students," says PSC's director of outreach and education, Cheryl Begandy, "and increases their willingness to stay with subjects they may otherwise find too complicated or just uninteresting." For Begandy and Pallavi Ishwad, education program director of PSC's National Resource for Biomedical Supercomputing (NRBSC), the goal is to help re-define high-school science instruction, so that it can better prepare future scientists, engineers and educators to participate in the cyber-savvy, 21st-century marketplace.

BEST STUDENT INTERNS AT TERAGRID 2011

Two Pittsburgh students, Annie Kayser (left) and Danielle Auth (right), with Pallavi Ishwad, were biomedical interns at the 2011 TeraGrid conference in Salt Lake City. Both took bioinformatics courses developed through BEST at Pittsburgh's Our Lady of the Sacred Heart High School, and both have entered college with majors, respectively, in biology and bioinformatics. After 2011 summer internships at PSC, they presented their projects, which involved the Python programming language, at the TG11 poster session.

BEST: Begun in 2007 by Ishwad, Better Educators of Science for Tomorrow (BEST) introduces high-school teachers to a bioinformatics curriculum adapted from an NRBSC program called MARC (Minority Access to Research Careers) for undergrad and graduate science students. Drafted and improved through classroom usage by an interdisciplinary group of STEM teachers, the BEST curriculum offers ready-to-use lesson plans for single-subject educators to extend their skills to the multidisciplinary outlook of bioinformatics, which draws on physics, chemistry, biology, computer science and math.

PSC's BEST summer workshops have introduced bioinformatics to six Pittsburgh area high-schools. In this year's workshop, from June 16 to July 22, PSC staff mentored 10 high-school teachers.

"You have provided a tremendous amount of expertise and guidance in helping to shape our program," said Edwina Kinchington, of the Pittsburgh Science & Technology Academy, one of six southwest Pennsylvania high schools that have adopted BEST curricula as part of permanent elective course offerings. PSC's Ishwad and Begandy both were recently renewed for a second three-year term on Occupational Advisory Committees for PS&TA, a public high-school with 65-percent minority enrollment. "The gift of this program to students is immeasurable," said biology teacher Rebecca Day of Frazier High School.

MORE INFO:
www.psc.edu/eot/k12/best.php

CAST: PSC this year received a $100,000 grant from the DSF Charitable Foundation that extends Computation and Science for Teachers (CAST), PSC's program — begun in 2008 — that has introduced many Southwest Pennsylvania STEM teachers to easy-to-use modeling and simulation tools for classroom learning.

The DSF grant funds a three-way effort among PSC and the Maryland Virtual High School Project, which helped to pioneer the use of computational thinking in high-school learning, and the Math & Science Collaborative of the Allegheny Intermediate Unit, which provides educational services to Allegheny County's 42 suburban school districts. Educators from these organizations are designing a professional development program for STEM teachers in western Pennsylvania to become leaders in integrating computational modeling and simulations into classroom learning.

"CAST," says PSC's Begandy, "brings to the classroom the same problem-solving, technology-rich approaches currently used in scientific research and in business."

CAST WORKSHOP FOR TEACHERS

At a July 25-27 workshop at PSC's Computer Training Center, southwest Pennsylvania STEM teachers piloted several CAST-developed modules in classroom teaching of modeling and simulation.

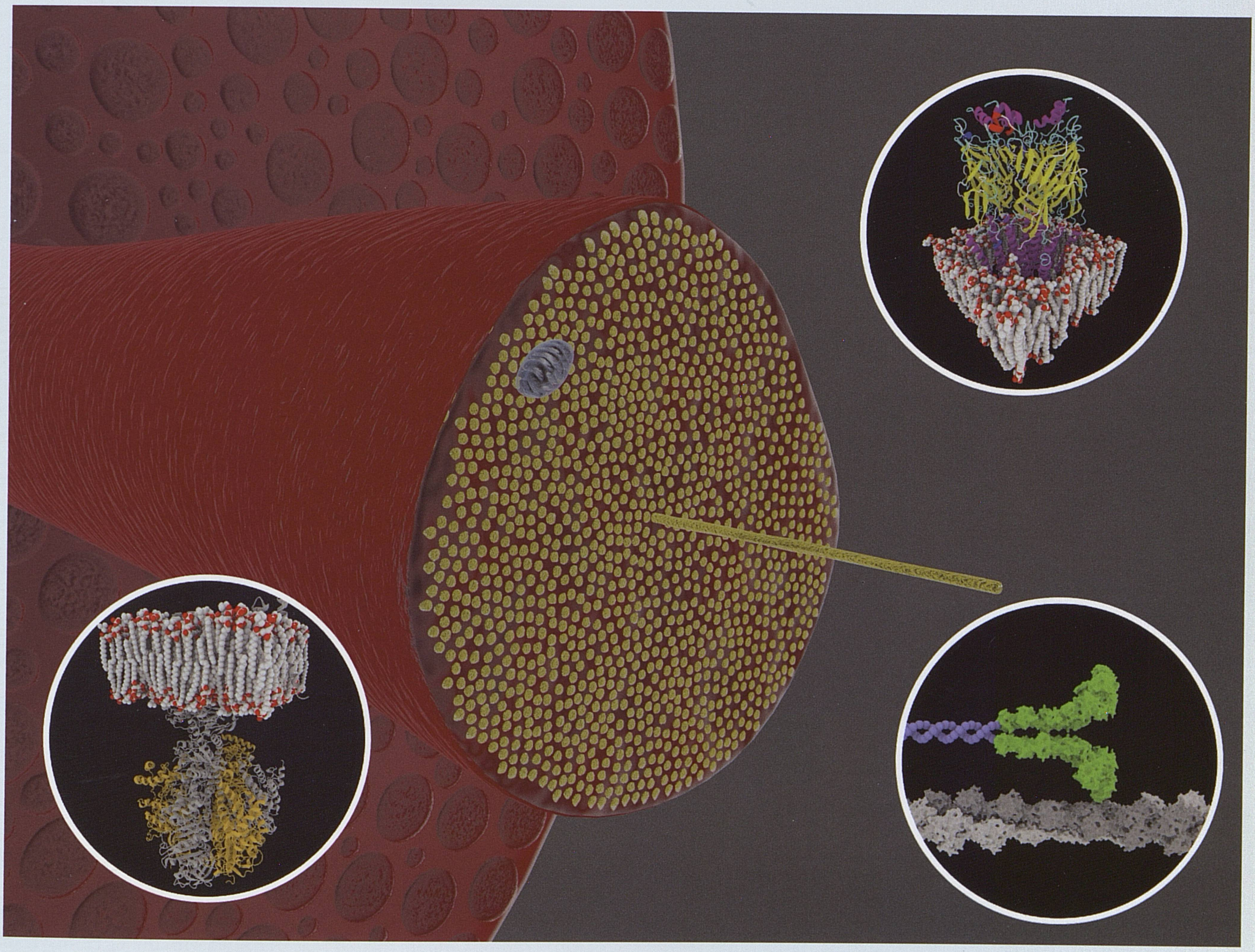
OPEN EDUCATION RESOURCES

Two of PSC educational programs, CMIST and SAFE-Net, provide open education resources on the World Wide Web for educators, students and parents. SAFE-Net's website provides free materials to help parents, educators, students and individuals understand questions of cyber-security associated with wide usage of the Internet.

Through NRBSC, PSC also provides modules and vivid 3D video animations developed through its CMIST program (Computational Modules in Science Teaching). Three CMIST modules are freely available through the website: Molecular Transport in Cells; Big Numbers in Small Spaces: Simulating Atoms, Molecules and Brownian Motion; and Enzyme Structure and Function.

SAFE-Net (free materials):
safenet.3rox.net

CMIST (free modules):
nrbsc.org/cmist



BIRTH OF A PROTEIN

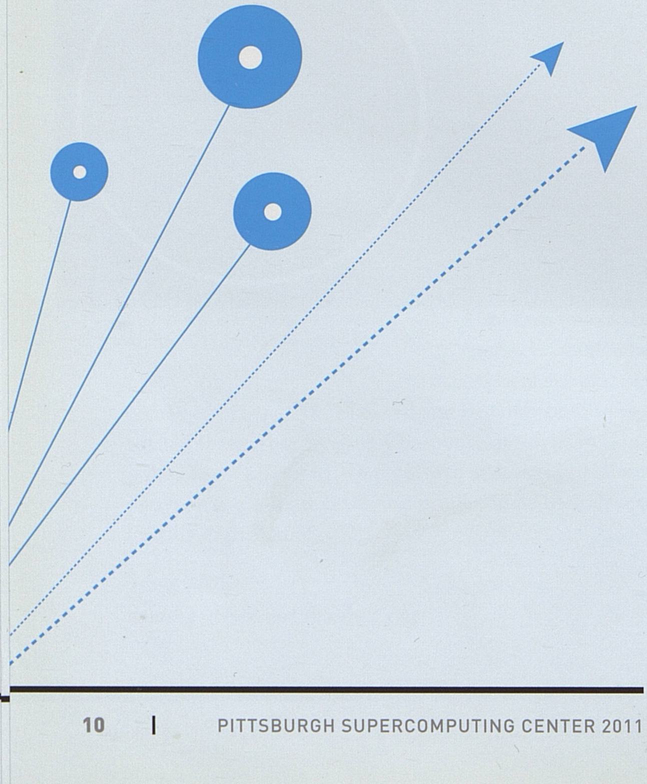
From the CMIST movie, Birth of a Protein, which uses vivid science animations to explain how proteins are created and generate energy for the body. The red cylindrical structure is a muscle fascicle, with an extending myofibril, composed of chains of two proteins, actin and myosin (inset: gray and purple with green heads). The other insets show ATP synthase (left) and acetylcholine (upper right), two molecules involved in protein creation and energy production in the mitochondria (gray coil).

NETWORKING THE FUTURE

One of the leading resources in the world for network know-how

PSC's Advanced Networking group is one of the leading resources in the world for knowledge about networking. Through 3ROX (Three Rivers Optical Exchange), a high-speed network hub, they operate and manage network infrastructure that connects many universities and schools in Pennsylvania and West Virginia to research and education networks, such as Internet2 and National LambdaRail, that link to universities, corporations and research agencies nationally. Their research on network performance and analysis — in previous projects such as Web100 and the NPAD diagnostic server — has created valuable tools for improving network performance. In a current project, Web10Gig, PSC network staff are helping to develop software to enable non-expert users to more fully exploit the bandwidth of advanced networks.

MORE INFO:
www.psc.edu/networking/



3ROX NEWS

Wendy Huntoon, PSC director of networking, has served in several national leadership roles in research and education networks and is currently chief architect in the office of the chief technology officer for Internet2, an advanced networking consortium led by the research and education community.

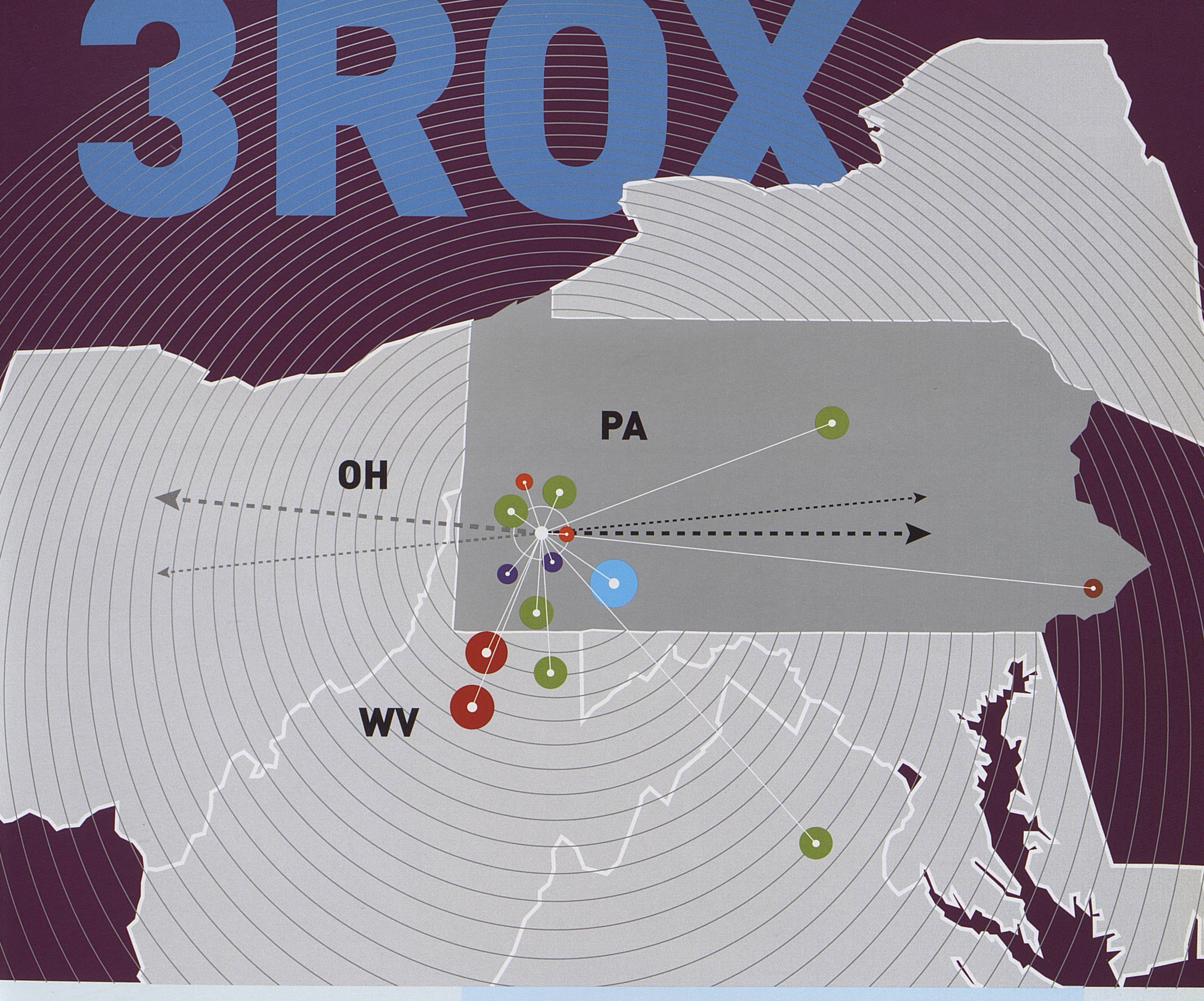
In January, 3ROX contracted with the National Oceanic and Atmospheric Administration for $2.6 million to provide a high-speed optical fiber link to NOAA's Environmental Security Computing Center (ESCC) in Fairmont, West Virginia. The connection to ESCC provides 10 Gigabit per second capability, and allows ESCC access to national and international research networks such as Internet2 and National LambdaRail.

In February, 3ROX partnered with Drexel University in Philadelphia to implement a five-fold upgrade to the Internet bandwidth of both 3ROX and Drexel at essentially no cost increase. Prior to partnering, 3ROX and Drexel each had individual one-gigabit (a billion bits per second) connections to Internet2, a high-performance research and education network that connects universities, corporations and research agencies nationally. Normally, the next level of service available would be 10 gigabits, which is cost prohibitive, but the new connection makes it possible to have two connections, each with five gigabits of committed bandwidth.

The partnership consolidates Internet2 connections in Pennsylvania from the previous three — 3ROX, Drexel and MAGPI (Mid-Atlantic Gigapop in Philadelphia for Internet2) — to two: MAGPI and 3ROX/Drexel. 3ROX serves universities, research sites and K-12 schools in western Pennsylvania and West Virginia, and Drexel connects the Drexel campus and its related research sites with the 14 Pennsylvania State System of Higher Education universities. With the new connection, both 3ROX and Drexel will be able to improve the quality and quantity of services they provide.

In June, 3ROX added Robert Morris University in Pittsburgh as a member, providing RMU with access to Internet2 and National LambdaRail and with improved network connectivity to other universities and area school districts.

3ROX



3ROX MEMBERS

● **UNIVERSITIES**
Carnegie Mellon University, Pennsylvania State University, Robert Morris University, University of Pittsburgh, Waynesburg University, West Virginia University, Norfolk State University

● **K-12 INSTITUTIONS**
Allegheny Intermediate Unit (AIU3), Arin Intermediate Unit (IU28), Beaver Valley Intermediate Unit (IU27), Intermediate Unit One, Northwest Tri-County Intermediate Unit (IU5), Riverview Intermediate Unit (IU6), City of Pittsburgh School District (IU2), Seneca Highlands (IU9), Central IU (IU10)

● **GOVERNMENT LABORATORIES AND FACILITIES**
The National Energy Technology Laboratory; NOAA Environmental Security Computing Center

● **BUSINESS**
Comcast, Westinghouse Electric Co.

● **RESEARCH NETWORK PARTNER**
Drexel University

● **OTHER**
Computer Emergency Response Team

NETWORK CONNECTIONS

➤ **NATIONAL RESEARCH NETWORKS**

Internet2 — 5 Gbps, ESnet — 1 Gbps, National LambdaRail PacketNet — 10 Gbps, XSEDE — 10 Gbps

➤ **NATIONAL COMMODITY INTERNET NETWORKS**

Global Crossing — 1 Gbps; Cogent — 1 Gbps

◄ **PITTSBURGH LOCAL EXCHANGE NETWORKS**

Comcast, MetNet & Cavalier

◄ **OTHER NETWORK CONNECTIONS**

Southern Crossroads (SOX) — 1Gbps, TransitRail-CPS — 4 Gbps, OARnet — 1 Gbps, FrameNet — 10 Gbps, WVNET — 10 Gbps

Note:
Gbps: a billion (Giga) bits per second.

THE NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

National Leadership in High-Performance Computing for Biomedical Research

Established in 1987, PSC's National Resource for Biomedical Supercomputing (NRBSC) was the first external biomedical supercomputing program funded by the National Institutes of Health (NIH). Along with core research at the interface of supercomputing and the life sciences, NRBSC scientists develop collaborations with biomedical researchers around the country, fostering exchange among experts in computational science and biomedicine and providing computational resources, outreach and training.

Joel Stiles (1958-2011), director of NRBSC (2005-2011).

The NRBSC team: (l to r) Christal Banks, Markus Dittrich, Nikolay Simakov, Boris Kaminsky, Hugh Nicholas, Pallavi Ishwad, Art Wetzel, Greg Hood, Troy Wymore, Jack Chang, Gary Blumenthal. (Not pictured: Jacob Czech)

ANTON PROGRAM EXTENDED

In September 2009, the National Institute of General Medical Sciences, part of NIH, awarded $2.7 million to NRBSC to support a partnership with D. E. Shaw Research, making an innovative new computing system, called Anton, available to U.S. biomedical scientists. This system, with hardware and software specialized for molecular dynamics (MD) simulations of biomolecular systems such as proteins and nucleic acids, runs MD up to 100 times faster than conventional supercomputers, making it possible to extend MD simulations into the millisecond range of biological time. Due to a successful initial year that included 47 biomedical projects (see pp. 18-23), NRBSC and DESRES extended the program, enabling a new round of projects to begin in October 2011.

"With this generous gift from D. E. Shaw Research and the funding provided by NIH," says NRBSC scientist Markus Dittrich, who coordinates the Anton program, "we are deploying a tool of unprecedented power for the benefit of biomedical researchers nationally."

MORE INFO:
www.nrbsc.org

NRBSC BIOMEDICAL COLLABORATIONS

Albert Einstein College of Medicine

Carnegie Mellon University

Duke University

Harvard University

Howard University

Grand Valley State University

The Salk Institute

University of California at Davis

University of California at San Diego

University of Pittsburgh

University of Pittsburgh School of Medicine

University of Puerto Rico, Medical Sciences Campus

University of Michigan

COMPUTATIONAL SERVICE & TRAINING

Since NRBSC's inception, PSC and NRBSC together have provided access to computing resources for more than 1,600 biomedical research projects involving more than 4,600 researchers at 285 research institutions in 46 states and two territories. Among these are several projects featured in this booklet (pp. 18, 24, 28 & 45). More than 4,800 researchers have participated in NRBSC workshops. NRBSC and PSC have also developed educational programs, CMIST and BEST (see pp. 8-9), that have provided training to high-school and undergrad students and educators in the Pittsburgh region and nationally.

PSC'S NRBSC WORKSHOPS

(2010-2011)

Hands-On Workshop on Computational Biophysics using NAMD and VMD

Summer Institute in Bioinformatics
(for minority-serving institutions)

Anton Training Workshop
(held three times)

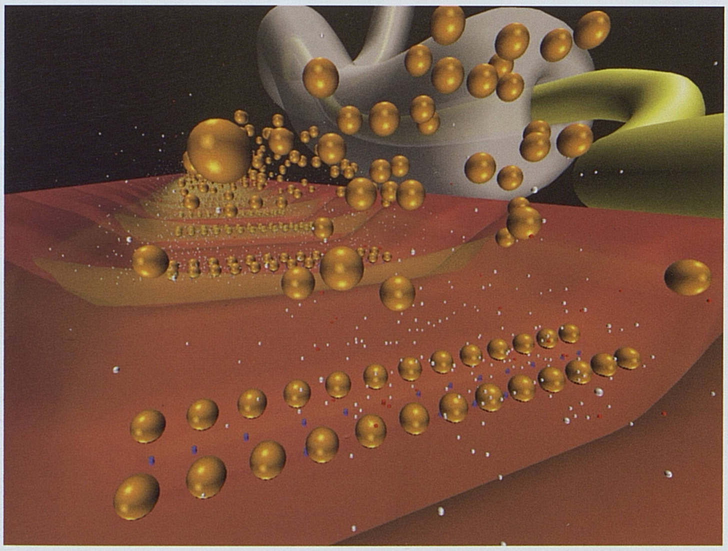
Bioinformatics Internship Program

Computational Modules in Science Teaching Workshop

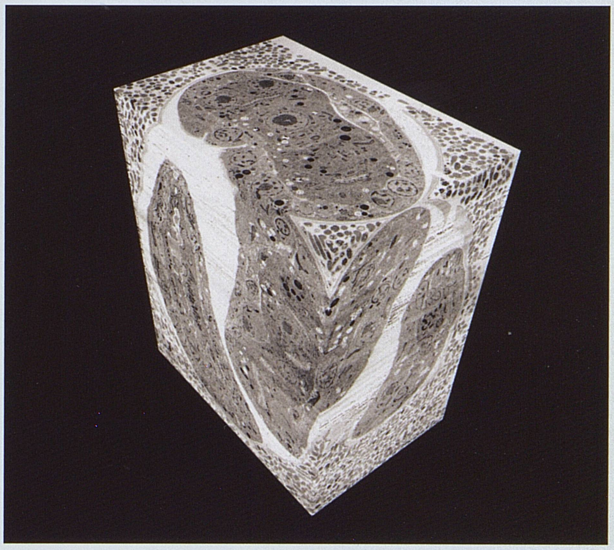
RESEARCH

NRBSC research focuses on three areas of biomedicine that span many scales of space and time: spatially realistic cell modeling, large-scale volumetric visualization and analysis, and computational structural biology.

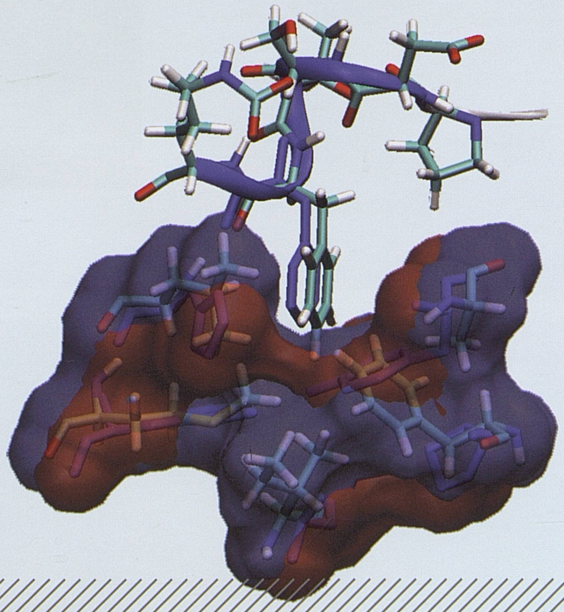
One of NRBSC's research highlights this year was a cover article in the March 10 issue of *Nature*, the prestigious international science journal (see pp. 24-27). As part of this project, NRBSC scientists processed and analyzed several tens of terabytes of electron microscopy image data.



Spatially realistic cell modeling centers on realistic 3-D simulations of movements and reactions of molecules within and between cells, to better understand physiological function and disease. MCell, DReAMM and PSC_DX software is developed at the NRBSC and used to model and visualize events such as (shown in this image) neurotransmission between nerve and muscle cells.



Volumetric visualization and analysis using NRBSC software enables multiple users to assemble and manipulate extremely large datasets and time series obtained from light and electron microscopy or CAT and MRI scans, etc. This cropped subvolume of a C. *elegans* embryo in its eggshell was assembled from 700 electron-microscopy images captured by Richard Fetter in Cornelia Bargmann's laboratory and aligned by Greg Hood at NRBSC. C. *elegans* is a roundworm much studied as a model organism.



NRBSC structural biology focuses on computational tools used to determine the structure of proteins from their amino acid sequence and development of quantum-mechanical simulation methods for biomolecules such as enzymes. This image shows co-varying residues of class D beta-lactamases, with the surface colored red for OXA-1 and blue for OXA-2. PSC-developed software enables researchers to simulate enzyme reactions, to reproduce experimental reaction rates and gain new insight into enzyme function, which facilitates design of new therapeutic drugs.

THE SUPER COMPUTING SCIENCE CONSORTIUM

Pennsylvania-West Virginia partners in development of clean power technologies

Formed in 1999 and supported by the U.S. Department of Energy, the Super Computing Science Consortium is a regional partnership of research and educational institutions in Pennsylvania and West Virginia. (SC)^2 provides intellectual leadership and advanced computing and communications resources to solve problems in energy and the environment and to stimulate regional high-technology development and education.

Through (SC)^2, Evergreene Technology Park in Greene County provides a resource that supports and encourages companies to collaborate with local universities in southwest Pennsylvania and West Virginia and to have access to PSC.

(SC)^2 co-chairs Lynn Layman, PSC (left) & Bob Romanowsky, NETL

Since the spring of 2000, a high-speed network — the first fiber-optic service to Morgantown, West Virginia — has linked the National Energy Technology Laboratory (NETL) campuses in Morgantown and Pittsburgh with PSC, facilitating NETL collaborations. Researchers at NETL and WVU have actively used this link to tap PSC computational resources.

PSC RESEARCH COLLABORATION

PSC has extended a research collaboration with NETL, begun during 2010, through the Regional University Alliance, which combines NETL's fossil-energy expertise with research at regional universities, including Carnegie Mellon. Through the end of 2010, PSC's work with NETL staff included implementation of VisIt, a software package for scalable visual analysis, on NETL's computing cluster and graphics accelerators. This work makes it possible for NETL researchers to use VisIt interactively with large data sets produced by MFIX (Multiphase Flow with Interphase Exchanges), NETL's award-winning software for simulating coal gasification and other clean-coal technologies.

"This work saves time for NETL engineers," says Nick Nystrom, PSC director of strategic applications, who coordinated the collaborative effort, "improves their analytical capabilities, and allows them to communicate results more effectively." The extended collaboration focuses on accelerating the processing speed of MFIX, along with analysis and visualization, through parallelizing the input/output.

(SC)^2 PARTNERS

National Energy Technology Laboratory

Pittsburgh Supercomputing Center

Carnegie Mellon University

University of Pittsburgh

Waynesburg University

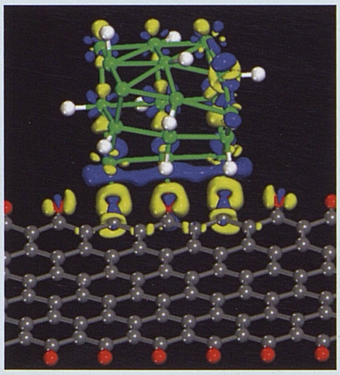
West Virginia University

MORE INFO:

www.sc-2.psc.edu

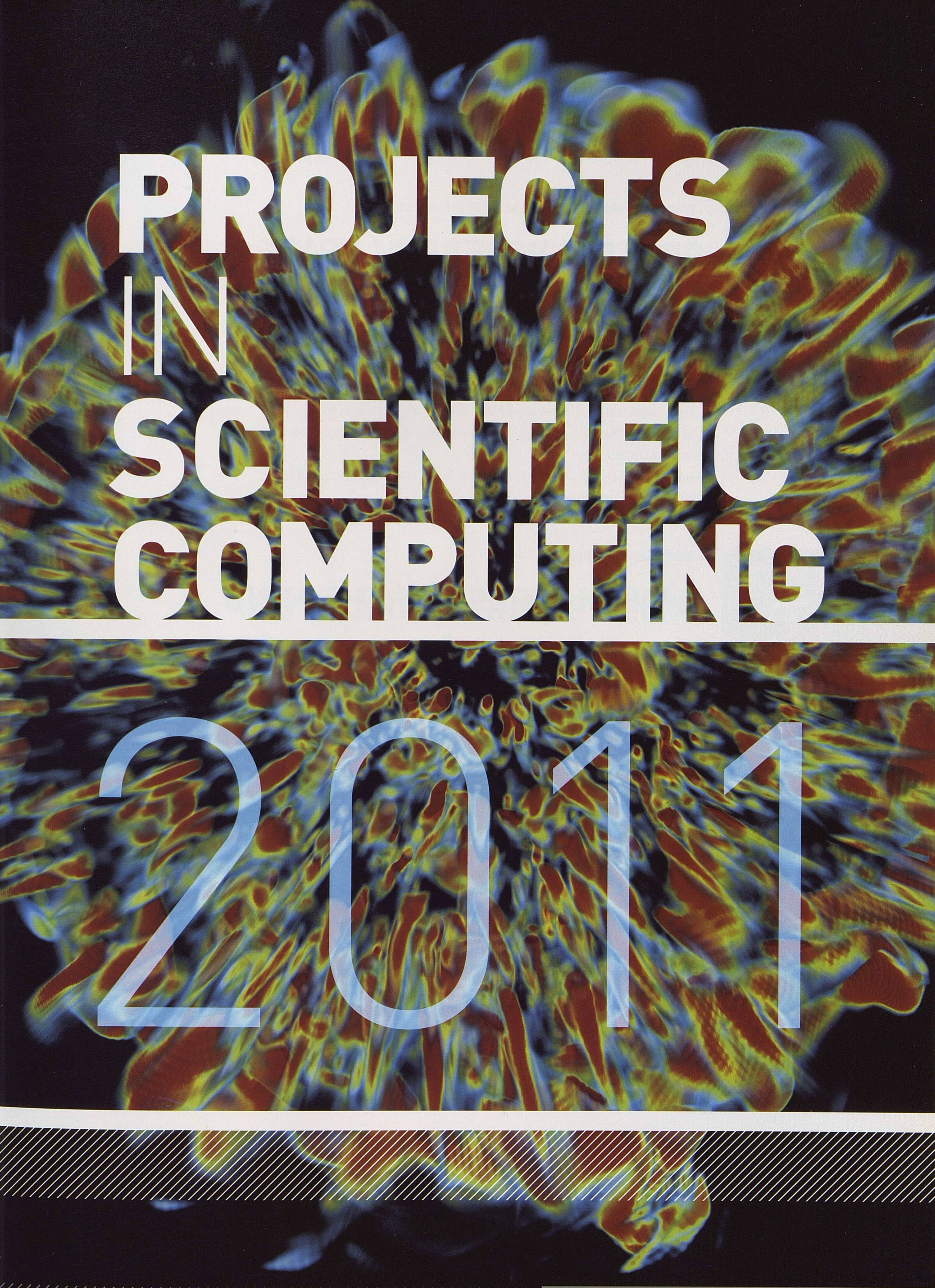
PSC & (SC)^2: RESEARCH FOR CLEAN ENERGY

Since the 1999 founding of (SC)^2, 52 (SC)^2 researchers have used PSC systems for a range of clean-energy related projects, including designs for advanced power turbines, fluidized-bed combustion, and a reactor to produce power from gasified coal. This work has used more than 6.5 million hours of computing time, over 370,000 hours during 2011.



HYDROGEN SENSITIVE NANORIBBONS OF GRAPHENE

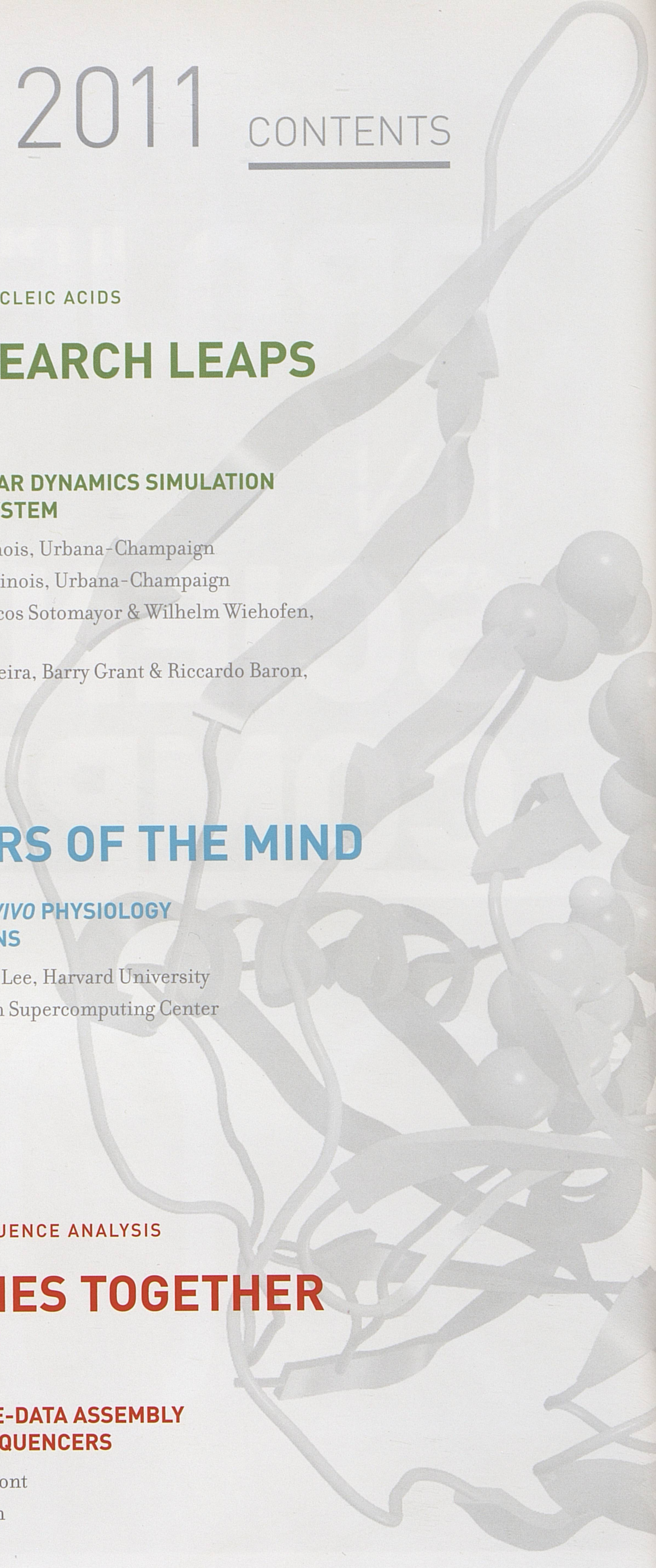
NETL scientist Dan Sorescu used Blacklight, PSC's newest system, to make progress on several projects involving quantum computations and simulations of nano-scale processes. One of these projects concerned potential applications of graphene, a one-atom thick layered form of carbon, as an electronic chemical sensor. Sorescu's calculations, combined with experimental work by Alexander Star's group at the University of Pittsburgh, suggest the feasibility of a technique to form interconnected graphene "nanoribbons" that can hold platinum nanoparticles so that the resulting structure exhibits a pronounced, selective electronic response toward hydrogen. The graphic (above) shows the variation of charge-density distribution (yellow & blue surfaces) at the graphene interface upon adsorption of hydrogen atoms (white) onto a cluster of 18 platinum atoms (green) located at the edge of a graphene nanoribbon functionalized with oxygen atoms (red).

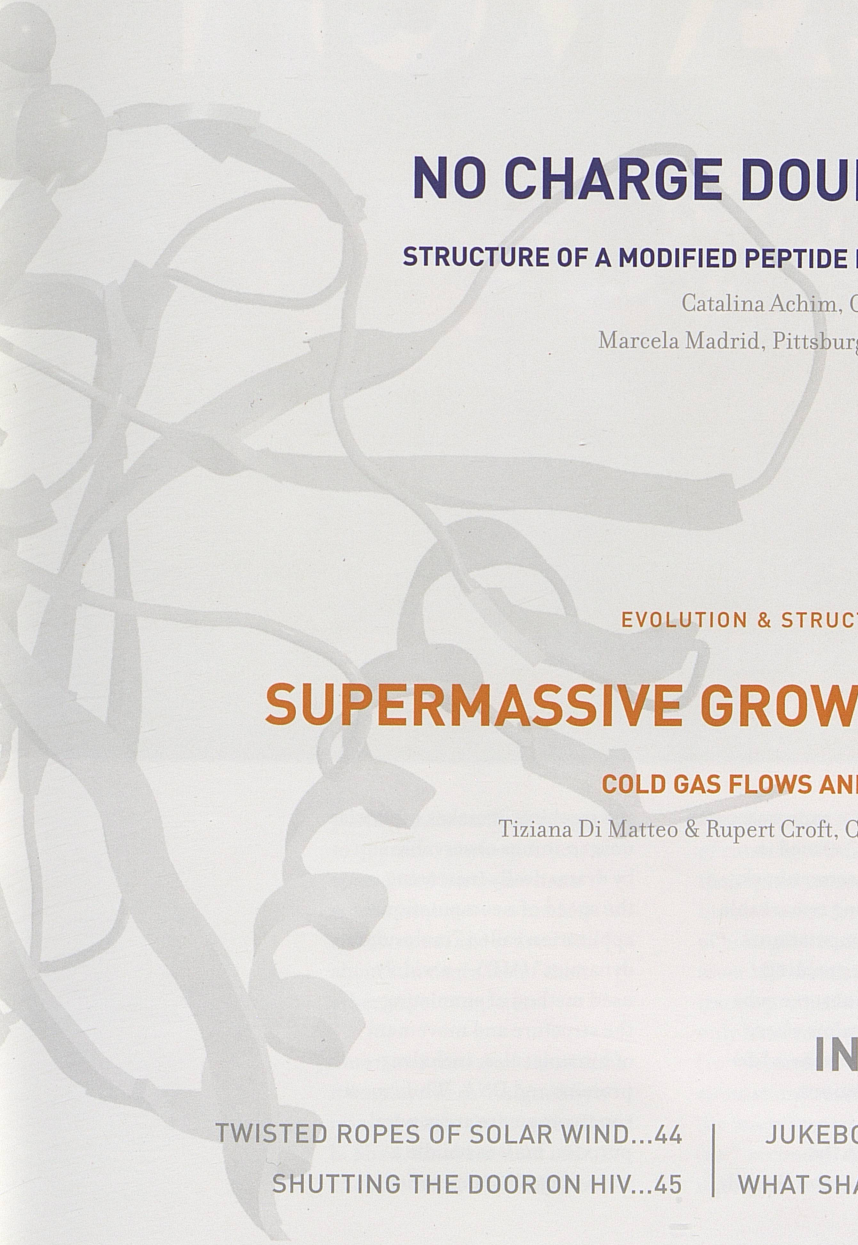


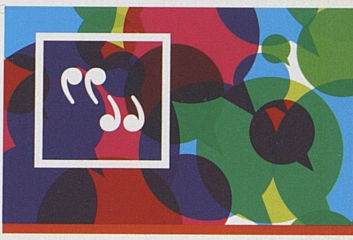
PROJECTS IN SCIENTIFIC COMPUTING

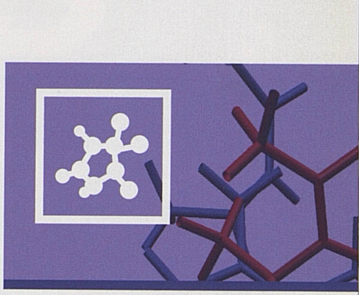
2011

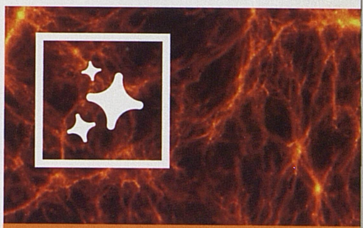
PROJECTS 2011 CONTENTS











IN PROGRESS

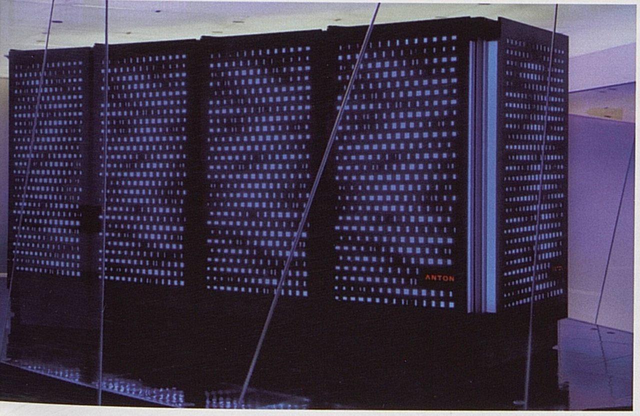
PROTEIN RESEARCH

Studies with Anton, a special-purpose supercomputer designed by D. E. Shaw Research and made available at PSC, have yielded new insights into the motion and function of proteins

In the 1670s in the small city of Delft in the Netherlands, a fabric merchant named Anton van Leeuwenhoek, who used magnifying lenses to count the threads in cloth, found ways to craft lenses of unprecedented power. His lens-making ability along with his natural curiosity led him to observe living things — bacteria, spermatozoa, blood flow in capillaries — no one had seen before. More than four centuries later, a supercomputer called "Anton," named in honor of van Leeuwenhoek, is demonstrating remarkable ability as a "computational microscope" — enabling researchers to observe sub-microscopic realms where proteins carry out their life-sustaining functions.

Although Anton the supercomputer can't literally see anything, it makes otherwise unseen things observable by dramatically increasing the speed of a computational application called "molecular dynamics" (MD) — a widely used method of simulating the structure and movement of biomolecules, including proteins and DNA. While most supercomputers are general purpose, built to handle a wide range of computational

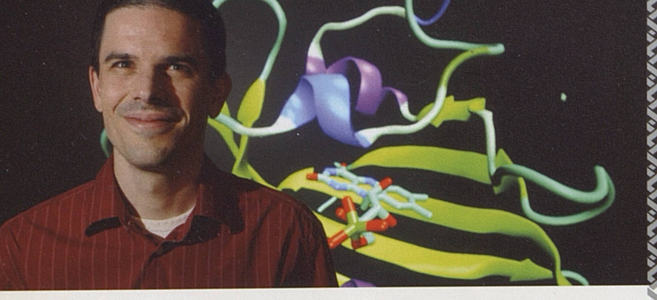
LEAPS FORWARD



ANTON
Research with Anton at PSC and at DESRES has demonstrated its ability as an unparalleled tool to advance understanding of biomolecules.

problems, from fluid dynamics to quantum physics, Anton's specialized hardware has a single purpose — to run MD simulations, and it does this about 100 times faster than other supercomputers.

Anton and the novel algorithms it employs were designed by a team of researchers led by David E. Shaw, chief scientist of D. E. Shaw Research (DESRES) in New York City. The objective of running MD faster, more than speed *per se*, is to simulate biomolecules for longer periods of biological time. Before Anton, most MD simulations could track a protein's movement for only hundreds of nanoseconds (10^{-9} seconds), with a few simulations reaching into the microsecond range (10^{-6} seconds). Anton makes it possible to routinely simulate biomolecules for tens to hundreds of microseconds, and in some cases into the millisecond range (10^{-3} seconds). At these longer timescales is typically where the most biologically important aspects of protein activity occur, which before Anton weren't accessible with MD simulation.



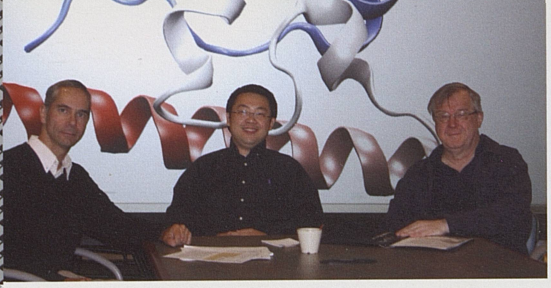
THE RESEARCHER

Markus Dittrich PSC

"Anton's ability to extend the timescale of molecular dynamics simulations," says Dittrich, who coordinates the Anton project at PSC, "has opened a new window on many important biological processes."

Since late 2008, DESRES has used Anton machines in its own internal research program, and in 2009 — in collaboration with the National Resource for Biomedical Supercomputing (NRBSC) at PSC — DESRES made one of its Anton systems available without cost for non-commercial research by scientists at universities and other not-for-profit institutions. PSC hosts this machine, supported by a two-year $2.7 million grant to NRBSC from The National Institute of General Medical Sciences (part of NIH).

In the summer of 2010, a panel convened by the National Research Council of the National Academies of Science reviewed proposals submitted by research groups from around the country and allocated time on Anton at NRBSC for 47 of these proposals. Many of these projects — as described in the rest of this article — have already produced new scientific insights about protein structure and function. Several papers have been accepted for publication, and other researchers have harvested unprecedented amounts of biomolecular simulation data that they're still analyzing. As a result of these successes, NRBSC and DESRES renewed the program, enabling a new round of projects that will begin in October 2011.



THE RESEARCHERS

Martin Gruebele (left), **Yanxin Liu** *(graduate student)* and **Klaus Schulten**, with the lambda-repressor fragment that they simulated in the background.

HOW PROTEINS GET IN SHAPE

"Anton is an amazing machine," says Martin Gruebele, of the University of Illinois, Urbana-Champaign (UIUC). Using Anton at PSC, graduate student Yanxin Liu, collaborating with Gruebele and with UIUC biophysicist Klaus Schulten, successfully simulated "protein folding" of a protein (lambda6-85) that, at 80 amino acids, is more than twice as large as the largest proteins whose folding had previously been successfully simulated and published.

The results are a major advance toward a challenge that scientists have called "the protein-folding problem." When a cell produces a brand-new protein, it's a droopy, unstructured chain of amino acids. Before it can begin to carry out its biological function, this chain must fold into the proper three-dimensional configuration. The result is a structured complex of folds, ribbons and helices, with clefts and notches that allow the protein to attach and release other molecules.

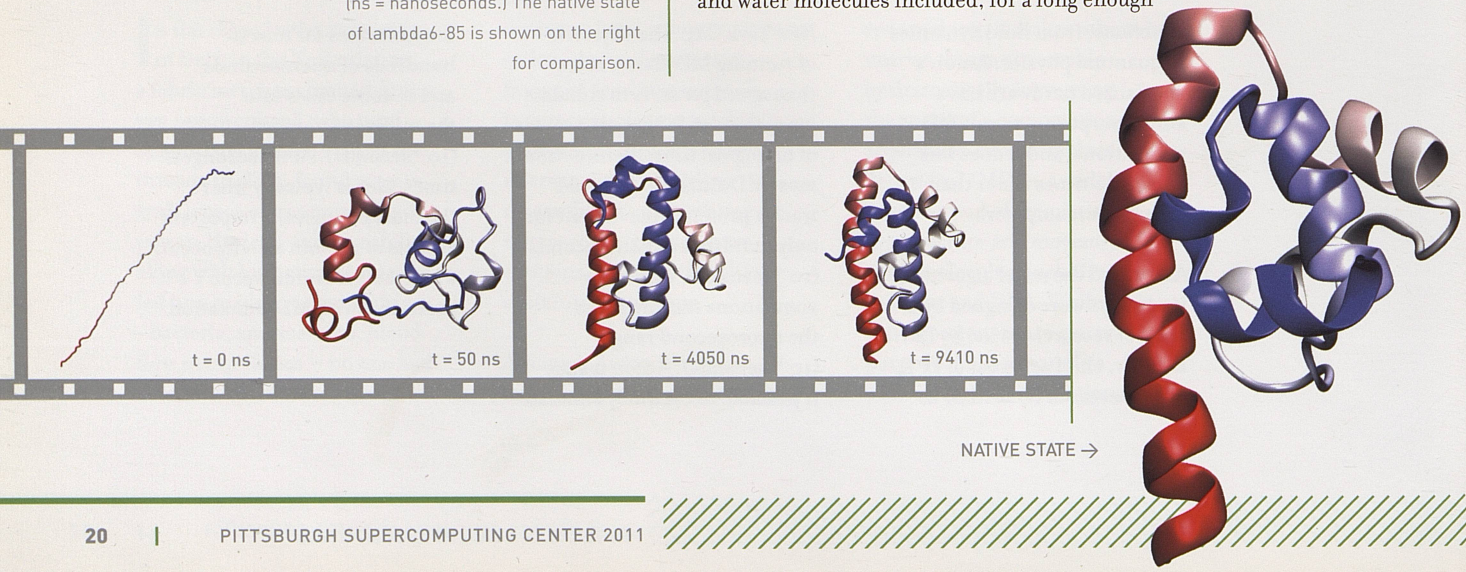
A major advance toward solution of the protein-folding problem

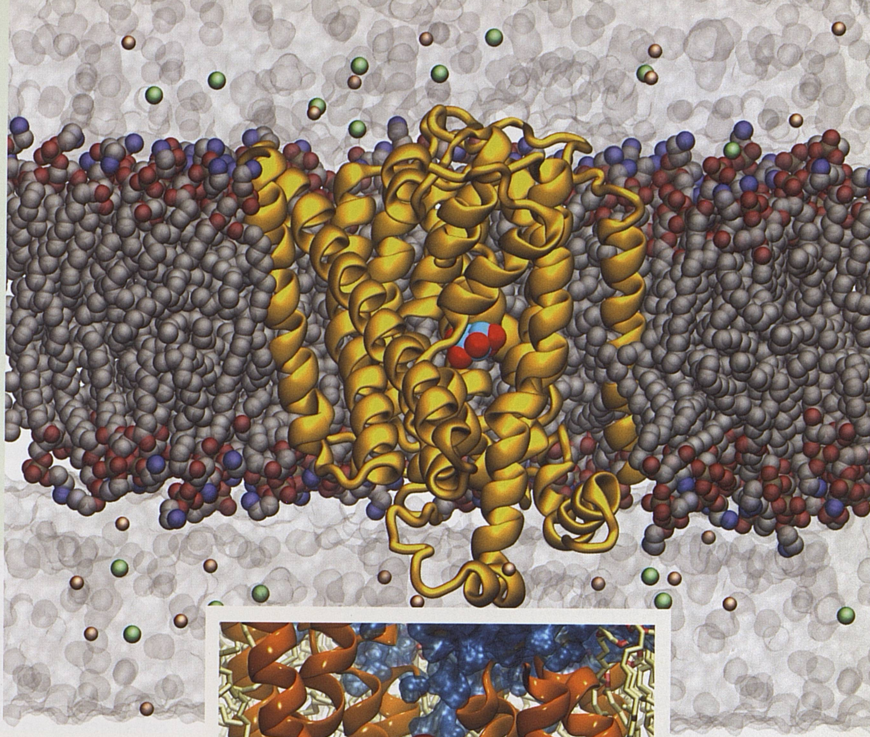
"The ability of a protein to fold into its characteristic three-dimensional structure," says Gruebele, "is crucial for living cells. Misfolded proteins not only lose their functions, but can also cause diseases, including Alzheimer's and Huntington's disease."

The Anton simulations of lambda6-85 by Liu, Gruebele and Schulten produced a folded form of the protein that compared well with experimental findings. "Anton enables simulations at full atomistic detail," says Gruebele, "all protein atoms and water molecules included, for a long enough

HOW A PROTEIN FOLDS

Snapshots of the folding of lambda6-85 during a simulation performed on Anton. (ns = nanoseconds.) The native state of lambda6-85 is shown on the right for comparison.





CELLULAR PASSAGEWAY

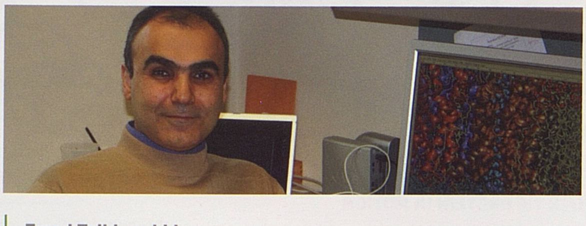
This image from the simulations shows a membrane transporter protein (gold) within the cell wall with water molecules (gray surface), sodium (pink) and chloride (green) ions inside the cell (top) and in the extracellular environment (bottom). A substrate molecule (galactose: red & blue) is midway in the transporter. Overall this simulation included 84,000 atoms.

time for a protein to be folded from 'first principles' on the computer."

For a protein of this size, this represents a big step toward the goal of being able to calculate the accurate folded structure of a protein from knowing only its amino-acid sequence. This is a major objective of molecular biology, since current methods of deriving a protein's folded structure depend on experimental processes — x-ray crystallography and NMR — that take months or years to find the structure of a single protein.

The "magic number," says Gruebele, is about 200 amino acids. "The largest single domain in most proteins is about that size, with much evidence that these domains fold relatively independently from each other. If we can do reliable simulations on that scale, we'll be able by running on the computer to do what otherwise takes experiments and years of analysis. Our work with Anton on lambda6-85 is exciting because it shows that this goal is within reach."

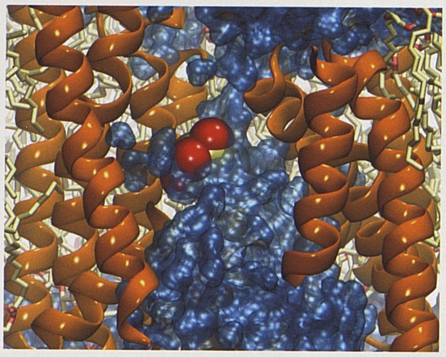
FINDING THE LEAK IN MEMBRANE TRANSPORTERS



THE RESEARCHER

Emad Tajkhorshid

Another Anton project, led by Emad Tajkhorshid also at UIUC, focuses on a family of proteins known as "membrane transporters." Like finely engineered doorway systems, these proteins reside in biological membranes and create highly regulated passageways for biomolecules — such as neurotransmitters — to cross from outside the cell to inside and vice-versa.



This closeup view from the simulation shows water molecules (blue surface) passing through the transporter along with the galactose substrate molecule.

With Anton, Tajkhorshid and collaborators simulated the structural changes in these transporters over a much longer period of time than has previously been possible. "Before Anton," he said, "we could simulate maybe 100 nanoseconds of protein motion. With Anton we were able to run several microseconds of simulation — more than 100 times longer in biological time."

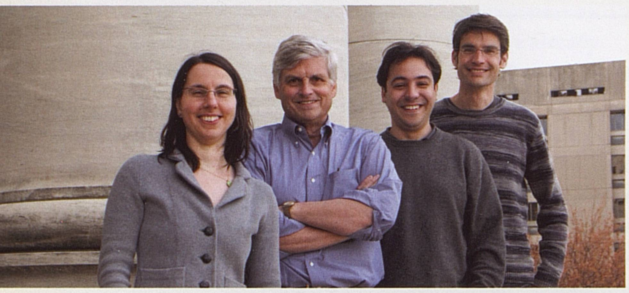
"With Anton we were able to run more than 100 times longer in biological time."

Because of how they work — one side must close as the other side opens — membrane transporters undergo large changes in structure as part of their function. This has made them difficult to study with MD. "These proteins," says Tajkhorshid, "are molecular machines that have to open on one side and close on the other in a highly coordinated manner as they go through their transport cycle. Large conformational change is key to understanding their mechanism, and with the MD simulations we could do before, we could observe initial motions only, not enough to be significant."

In being able to extend transporter simulations into the microsecond range, Tajkhorshid has begun to characterize a phenomenon involving water "leaks" through these proteins. During their function, as they switch from one side to the other being open, they take in water molecules that can then pass through to the other side along with the "substrate" molecule. This had been noticed experimentally, says Tajkhorshid, but not understood in detail.

Tajkhorshid's Anton simulations show water leakage in four different cases. "We have a collection of simulations," he says, "in which we've observed this phenomenon for all the transporter sub-families that we investigated." Their findings, which they have reported in a submitted paper, point toward further experiments and suggest that current understanding of membrane transporters may need to be revised.

INNER-EAR PROTEINS



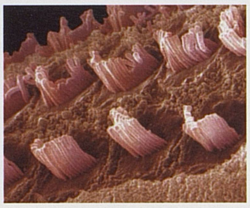
THE RESEARCHERS

Rachelle Gaudet, **David Corey**, **Marcos Sotomayor** and **Wilhelm Wiehofen**

Sotomayor credits PSC staff for the group's success with Anton: "It's a new machine and very powerful but you have to learn how to use it. PSC put on a workshop that was very helpful in getting these simulations started."

A team of researchers at Harvard used Anton to arrive at better understanding of proteins that make it possible to hear. Residing at the tips of hair cells in the inner ear, these proteins form a thin filament called the "hair-cell tip link," which is directly involved in transforming mechanical vibrations into the sensation of sound.

Hair bundles on the surface of the hair cells are comprised of "stereocilia" — flexible, finger-like structures arranged in rows. When sound vibration stimulates the hair-cell membrane, it stirs the stereocilia to movement that's similar to a field of grain stirred by wind. The tip links connect each stereocilia to its neighbor and, as they stretch, convey the force of this wave-like movement to ion channels at the stereocilium tip, which then open to trigger electro-chemical signals to the brain.



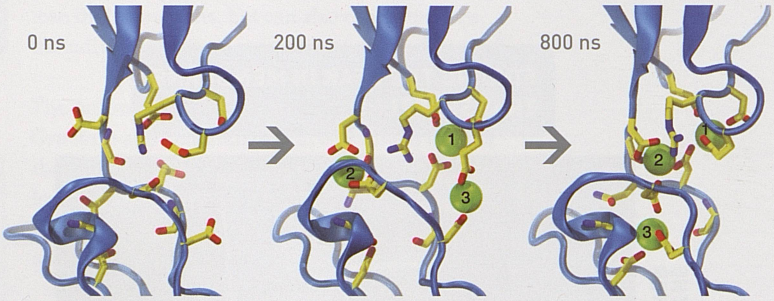
Credit: Fred E. Hossler, inner-ear hair cell of a guinea pig

GOOD VIBRATIONS

Electron microscope image of hair bundles on the surface of an inner-ear hair cell. The hair-like stereocilia respond to the force and pitch of vibration with wave-like motions that trigger corresponding electrochemical signals to the brain. Stereocilia are approximately one to three microns (millionths of a meter) long.

Professors David Corey and Rachelle Gaudet and post-doctoral fellows Marcos Sotomayor and Wilhelm A. Weihofen collaborated on the project. Laboratory work to find the 3D crystallographic structure of the tip link — a combination of two proteins — set the stage for simulations using Anton. By extending MD into the microsecond range, this work gives a much more complete picture of how one of these tip-link proteins (cadherin-23) changes when it binds with calcium ions.

Research has shown that genetic defects in cadherin-23 cause deafness, and these mutations affect the part of the protein that binds with calcium. "We know that mutations target the calcium-binding site," says Sotomayor, "and now with the simulation for the first time we can see the whole process of calcium binding, to clarify why some amino acids are important and might be related to deafness."

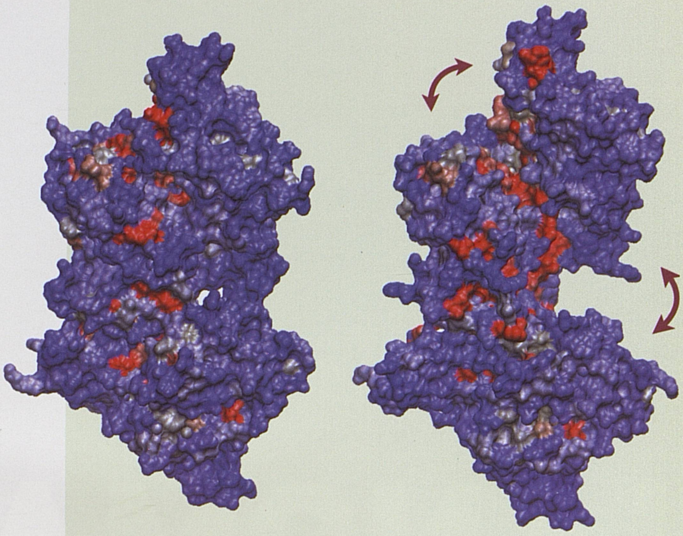


HAIR TIP PROTEIN IN MOTION

Snapshots from Anton simulations of calcium ions (numbered green spheres) binding to cadherin-23 (blue, binding site in colored stick), a protein essential to hearing. The first frame (left) shows the protein alone; the second (200 nanoseconds of biological time) shows calcium beginning to bind; the third (800 ns) shows a calcium-bound structure that agrees well with the crystallographic structure.

"For the first time we can clarify why some amino acids might be related to deafness."

One of their Anton simulations started with the protein and calcium not bound to each other and evolved to a structure, with calcium bound, that matches well with the previously obtained 3D crystallographic structure (which includes calcium). The researchers are now looking more closely at the overall calcium-binding dynamics, in particular which amino acids are involved, which could lead to better understanding of deafness. "We expect that those amino acids," says Sotomayor, "are involved in hereditary deafness."



EXPLORING THE OPEN STATE
The closed (left) and open state of the TcPR enzyme, with arrows indicating changes during simulation of the open state.

ATTACKING CHAGAS DISEASE

THE RESEARCHERS
[from top to bottom]
César de Oliveira
Andrew McCammon
Barry Grant
Riccardo Baron

Chagas disease threatens millions of people from the southern United States to Argentina. It's caused by a protozoan parasite, *Trypanosoma cruzi*, that enters the blood, usually by way of bites from a common insect, the triatomine bug. Anti-parasitic medications, which can have fairly severe side-effects, are sometimes helpful, but 20 to 40-percent of people chronically infected with Chagas develop life-threatening heart and digestive system disorders.

With the aim of finding more effective drug therapies, Andrew McCammon of the University of California, San Diego and post-doctoral fellows César de Oliveira, Barry Grant and Riccardo Baron used Anton to simulate an enzyme from the Chagas parasite — *T. cruzi proline racemase* (TcPR). Research has shown that TcPR triggers the parasite's ability to, in effect, trick the immune system and sustain itself as an invader in the blood stream. Their simulations, which extended to three microseconds for two different forms of the enzyme, reveal new information about how TcPR changes its structure, and offer new insight for designing a drug to defeat the disease.

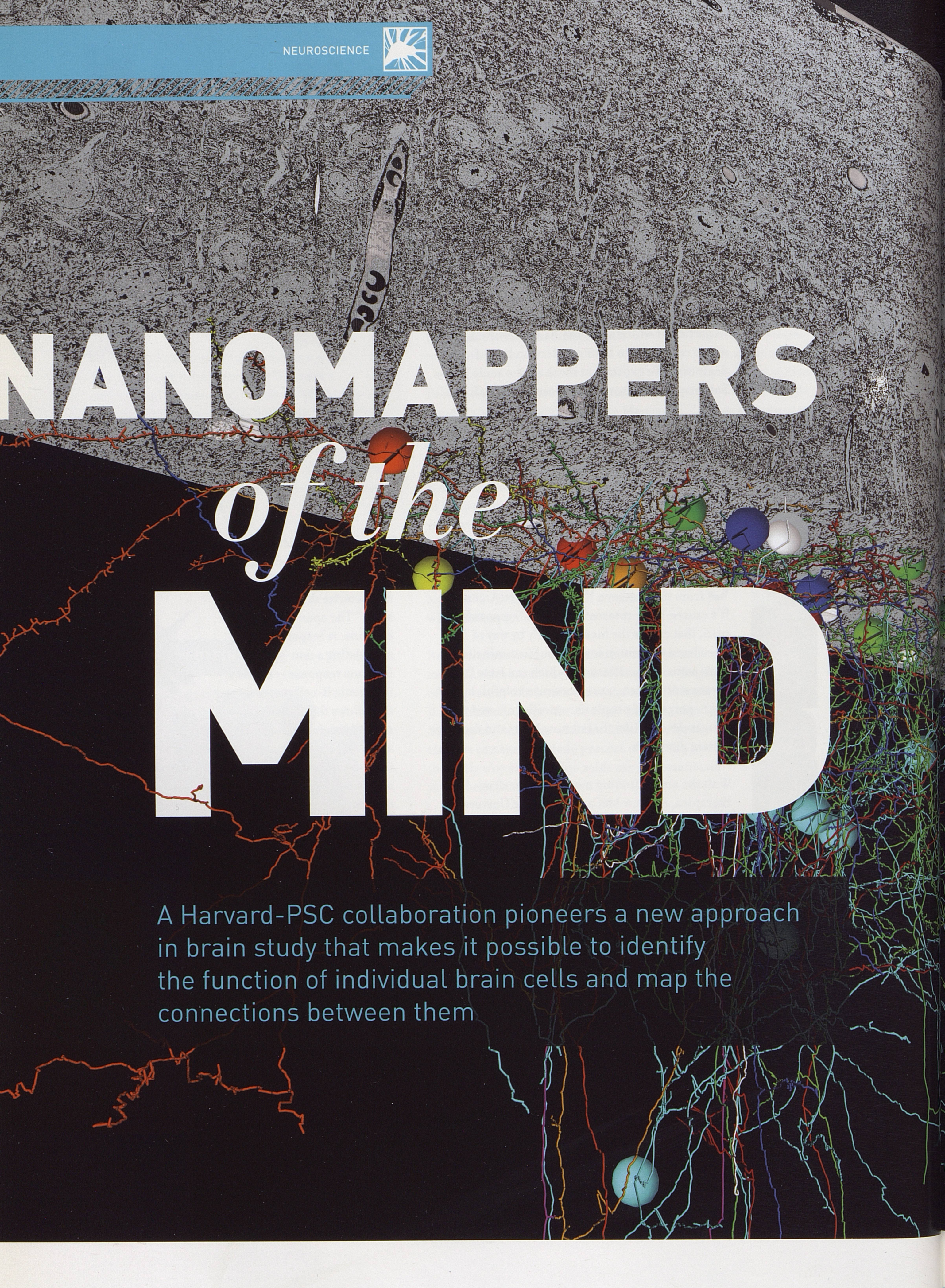
Their simulations offer new insight for designing a drug to defeat Chagas disease.

TcPR has two major forms: closed, when bound with a "ligand" — for which a 3D structure is available, and open, when it's free in solution. "No one has had access experimentally to the open state," says de Oliveira. "We've been doing MD to characterize this unexplored state." The open state, he explains, is responsible for stimulating a non-specific immune response — called a mitogenic B-cell response — that allows the parasite to establish infection and avoid a specific immune response that otherwise might protect the host organism.

Triatomine bug (*Rhodnius prolixus*, subfamily *Triatominae*), aka the assassin bug, cone-nosed bug and kissing bug, depending on environs, one of several related species of nocturnal blood-sucking insect that transmit Chagas disease.

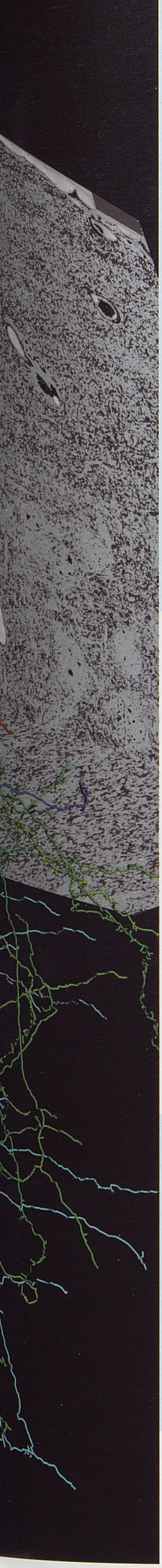
Their simulations with Anton show, for the first time, the beginning of the opening of TcPR, with several residues around the active sites becoming exposed to solvent. "You can track what protein segments are involved in the opening motion," says de Oliveira. With this information, the researchers are now collaborating with an experimental group that will test to see if the segments identified in the simulation are involved in eliciting immune response. In further work, McCammon, de Oliveira, Grant and Baron anticipate using the simulation data from Anton to computationally screen thousands of potential "inhibitor" molecules, to find compounds that can block TcPR's active site from triggering a mitogenic response and thereby potentially lead to a drug that can defeat the parasite's ability to threaten life.

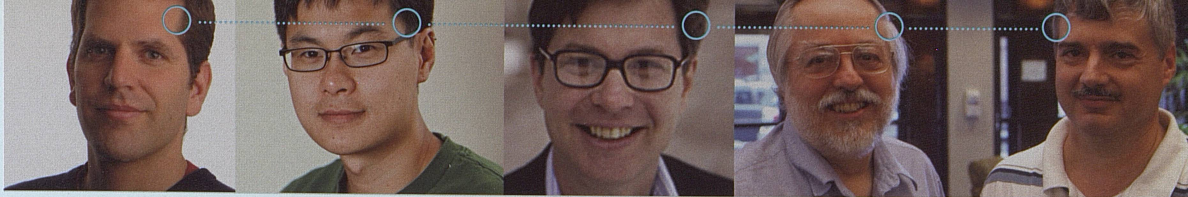
MORE INFO:
www.psc.edu/science/2011/antonhighlights



NANOMAPPERS *of the* MIND

A Harvard-PSC collaboration pioneers a new approach in brain study that makes it possible to identify the function of individual brain cells and map the connections between them





THE RESEARCHERS

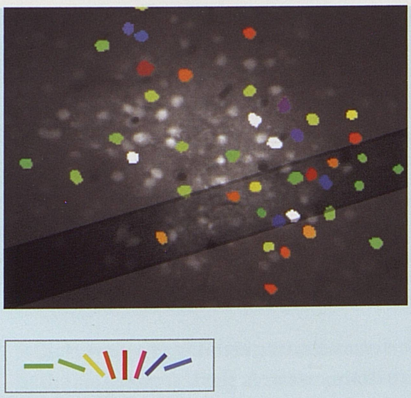
[from left to right]

Davi Bock HOWARD HUGHES MEDICAL INSTITUTE
Wei-Chung Allen Lee HARVARD UNIVERSITY
Clay Reid HARVARD UNIVERSITY
Art Wetzel PITTSBURGH SUPERCOMPUTING CENTER
Greg Hood PITTSBURGH SUPERCOMPUTING CENTER

March 10, 2011: "Untangling Neural Nets" said the big-print headline on the cover of *Nature*, the international journal of science, with a colorful image from work by scientists at Harvard and the Pittsburgh Supercomputing Center. According to a news comment in the same issue, their paper reported "an exciting and pioneering approach . . ." by which they "achieved a new feat . . . a way of directly studying the relationship of a neuron's function to its connections."

The research described in these glowing terms culminates several years of work at Harvard as well as prodigious data transmission, management and image processing at PSC. The result is a notable first step toward a major goal of neuroscience: a wiring diagram of the brain. "We've just begun to scratch the surface," says Clay Reid, professor of neurobiology at the Harvard Medical School and Center for Brain Science, who led the project, "but we're moving toward a complete physiology and anatomy of cortical circuits."

Their study relied on a series of innovations with advanced technology, beginning with experiments that identified individual brain cells — known as *neurons* — of a mouse as they respond to what the mouse is seeing. "The first part of this work," says Reid, "is something we've been doing for five or six years — literally watching the brain see, with neurons reacting to very specific elements in the field of view."



NEURON FUNCTION

From imaging of a mouse visual cortex, this image shows individual neurons color-coded according to orientation of visual stimuli to which they respond.

The next challenge involved capturing high-resolution images of the extremely small volume of brain involved — about the size of the period at the end of this sentence. Within this tiny patch of the visual cortex, the part of the brain that processes impulses from the retina, they had identified 100 neurons according to function. With very high-resolution imaging, they aimed to make sense from the tangled, tentacle-like mass of neural structure and to relate neural wiring to function. "The big challenge," says Reid, "is tracing the connections between neurons."

To get the high-resolution images needed, Reid and Ph.D. student Davi Bock (now head of a laboratory at Janelia Farm, the research campus of the Howard Hughes Medical Institute) and postdoctoral fellow Wei-Chung Lee developed a souped-up version of transmission electron microscopy (TEM). They prepared the brain tissue and — with a precision diamond knife called an "ultramicrotome" — cut ultra-thin slices (40 nanometers, the thickness of a few hundred atoms) of part of the volume that included 10 of the functionally identified neurons. With a specialized TEM camera array, they imaged these sections.

The imaging presented a massive data-processing task. In late 2007 a fortuitous meeting of minds (pun intended) occurred at a brain science conference. Discussions between Reid, Bock and bio-imaging expert Art Wetzel of PSC's National Resource for Biomedical Supercomputing (NRBSC) opened the door to a way forward in handling the TEM data. By April 2009 Wetzel and his NRBSC colleague Greg Hood had tools in place at PSC to receive from Harvard, store and process more than a terabyte (a trillion bytes, equivalent to 1000 copies of the full set of Encyclopedia Britannica) of TEM data per day.

Over a six-month run, April to September, the NRBSC scientists transmitted more than 110 terabytes of data and collected more than three-million TEM camera frames from Harvard. PSC network staff worked closely with Harvard to maximize bandwidth performance. "This was near the limits," says Wetzel, "of what could be sent using commodity best-effort network service."

Wetzel and Hood archived the image data (using PSC's file archival system), and at the same time, with a workstation custom-built for this job, began the task, a combination of art and science, of digitally stitching frames into sections (as many as 14,000 frames per section) and then stacking these quilt-like sections to recreate the imaged brain volume for 3D viewing on a computer screen. In 2010, after months of work at NRBSC, the Harvard team used this 3D model to manually trace the axons (output wires) from each of the functionally identified neurons to its junction (synapse) with a dendrite (input wire) of another neuron — and beyond, to the boundaries of the imaged volume.

Reid and his colleagues, in effect, crawled through the brain's dense thicket, neuron to neuron, and mapped a small part of the visual cortex. "This gives us a new approach," says Reid, "to answer the question, 'How does the brain see?' We can finally look at circuits in the brain in all of their complexity. How the mind works is one of the greatest mysteries in nature, and this presents a new and powerful way to explore that mystery."

QUILT PATCHING & SLICE STACKING

Research on the visual cortex over nearly a century has shown that it is organized into circuits according to visual function. Neurons that respond to vertical features in the field of view — trees or telephone poles, for instance — are interspersed in the mouse visual cortex with neurons that respond to horizontal features. The objective of this research is to reverse engineer the wiring in order to get at deeper understanding of how a circuit works, in particular to understand how neurons that respond to different features make connections in a local circuit. "To understand the cerebral cortex," says Reid, "we'd like to go one circuit at a time. A circuit is roughly 10,000 neurons with tens of millions of connections between the neurons."

Ten neurons, the researchers are acutely aware, is barely a start, but a start, and they're busy planning an expanded study. "By historical standards, this was a large volume," says Bock, for whom this work constituted his Ph.D. dissertation, "but it was barely big enough to contain some interesting cortical circuitry."

"What we've done," says Wetzel, referring to the paper in *Nature*, "is about 1/80th of the target volume for our next step, a cubic millimeter, large enough to encompass a circuit." In preparation for the larger volume, he and Hood have begun upscaling their storage and processing capabilities to handle as much as 100 terabytes, and expect to be prepared to handle data transmission at the scale of petabytes (1000 terabytes) in two to three years.

To get an idea of the quantity of information-capture involved, imagine the brain as a wedge of cheese. If each TEM-prepped section were a millimeter thick, roughly a thin slice of cheese (instead of 40 nanometers), and the lateral dimensions increased proportionately, the cheese slices would be larger than a basketball court. A cubic millimeter of brain will yield 25,000 of these basketball-court sized cheese slices.

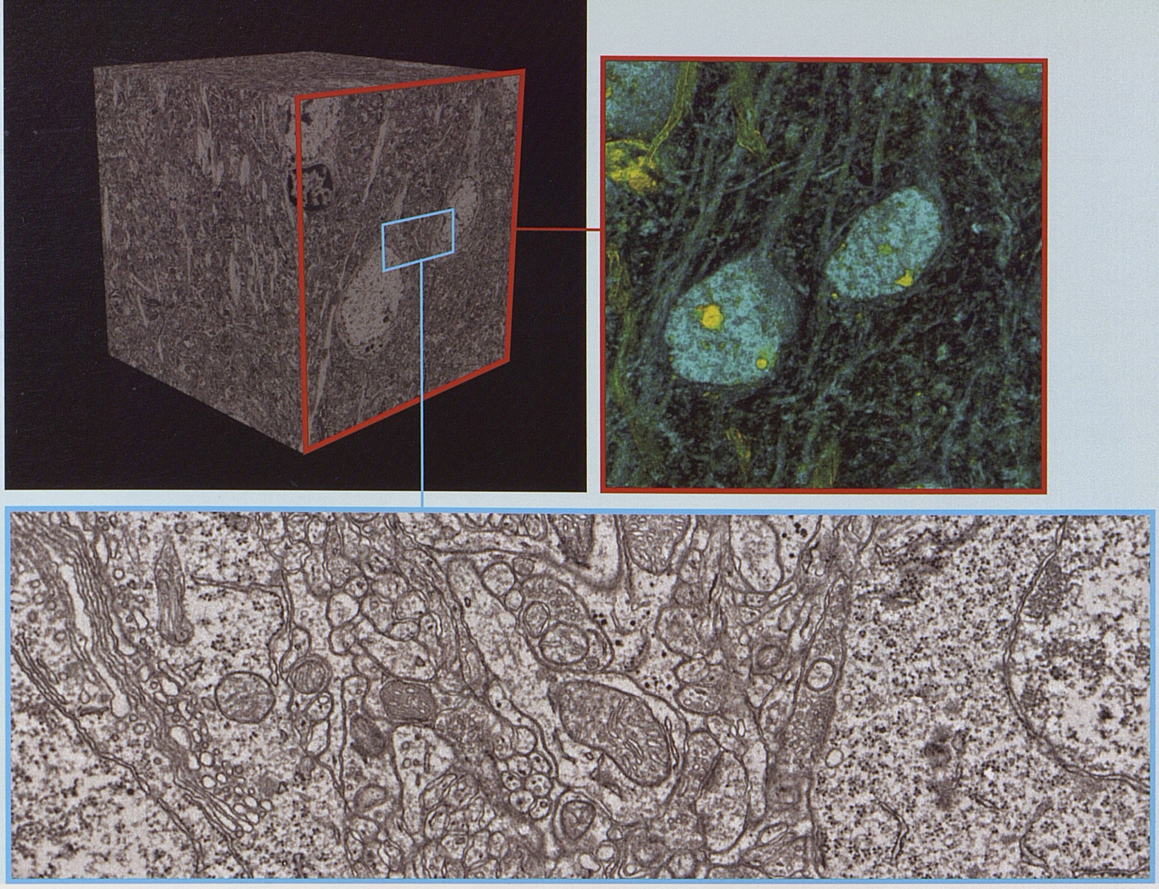
For the automated stitching of frames and post-processing of the immense TEM data sets, Hood and Wetzel applied software methods they developed in earlier volumetric imaging, mainly with the roundworm (C. *elegans*), adapting them to the Harvard data and, in some cases, creating *ad hoc* methods. The latter included adjusting for distortions in the frames that occur as part of TEM imaging. "Some sections," says Hood, "had pronounced shear distortion. Since this is very regular, we could mathematically compensate for it before addressing the irregular distortions."

To stitch the individual frames, intentionally imaged with overlap, into a single mosaic, they use various search methods (including fast Fourier transform correlations) to match information in adjacent frames. This process, says Wetzel, matches frames both spatially and in intensity to produce a nearly seamless image of each section.

To map each section to its neighboring sections, they apply a "pair-wise" registration algorithm, compensating for deformations that inevitably occur when cutting tissue so thinly. They next construct a "spring model" of the entire stack of sections, with the pair-wise registration maps guiding the placement of springs between adjacent sections. Finally, by letting this spring model relax, they obtain a 3D alignment of the stack and can produce a finished volume for viewing and analysis.

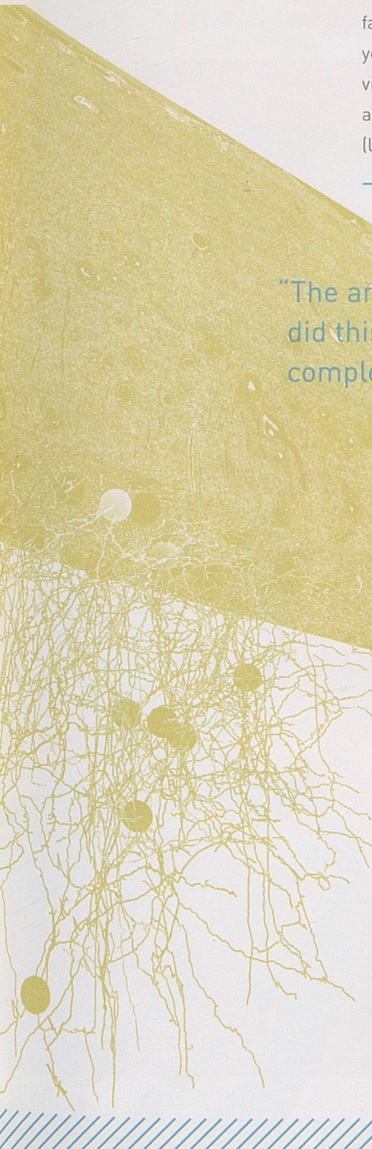
TIP OF THE ICEBERG

Along with demonstrating viability of a powerful approach to brain research, the Harvard researchers also produced new evidence tending to confirm earlier studies about "inhibitory neurons" — neurons that, rather than transmitting an excitatory electro-chemical pulse, suppress the activity of other neurons. Tracing the axon-to-dendrite connections within the imaged volume



ZOOMING-IN ON THE BRAIN

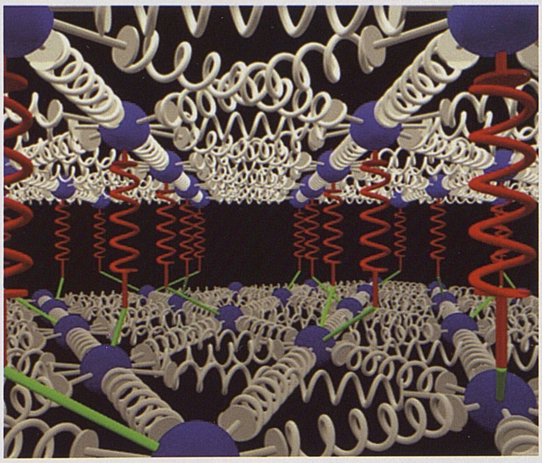
A cube of the EM-imaged volume with one face in two dimensions (colored by electron density, yellow: dense to aqua: less dense) and a zoomed-in view (blue rectangle) illustrating the density of axons and dendrites between two cell bodies (left and right sides of the rectangle).



> "The amount of data they took-in and did this very precise alignment with was completely unprecedented."

showed that the inhibitory neurons, which the researchers could identify by their structure, had no functional preference; connections to their dendrites arrive from the functionally identified neurons without regard to the visual response properties of the transmitting neuron. Understanding these relationships can be important, says Reid, because many neurological conditions, such as epilepsy, seem to be the result of neural inhibition gone awry.

Still, the accomplishment, in this case, lies less in the scientific finding than the proof of method, for which Reid gives large credit to the partnership with PSC. "The amount of data they took-in and did this very precise alignment with," he says, "was completely unprecedented. We couldn't have done our science unless we had this team to wrestle the large dataset to the ground."

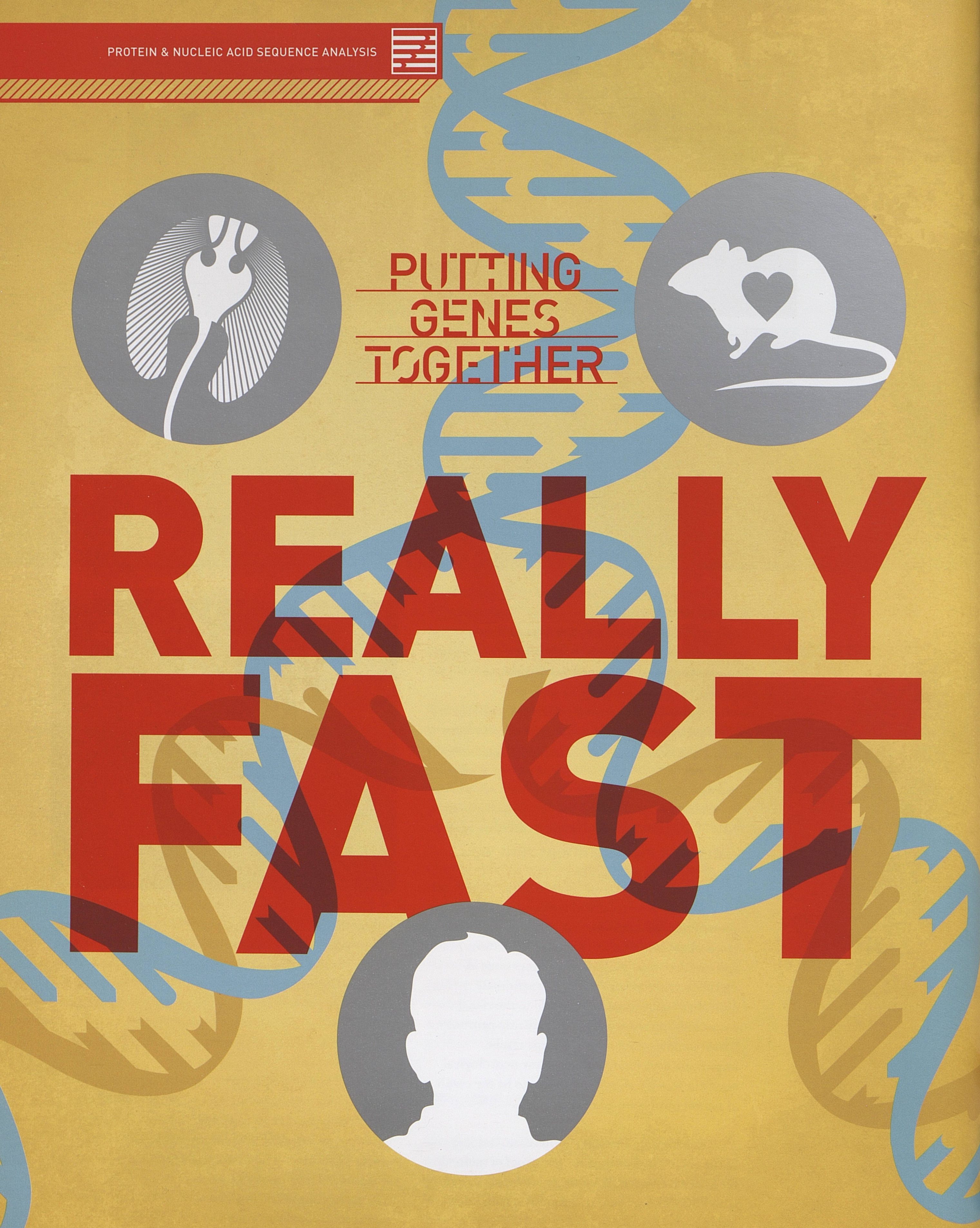


SPRINGING TO ALIGNMENT

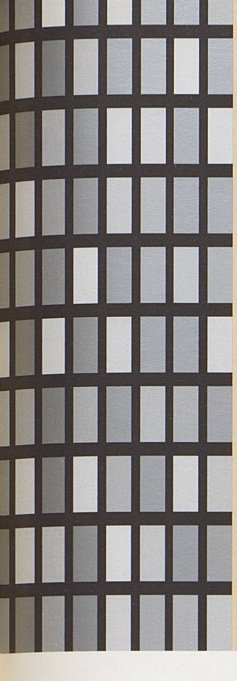
Blue spheres represent imaginary nodes placed every 64 pixels in the sections. The white springs within a section oppose distortion. The red springs represent the pair-wise mapping between adjacent sections, with the green offsets representing the displacements necessary to bring the sections into good alignment.

As Wetzel and Hood prepare to handle more data, Reid and his colleagues are scaling-up their TEM platform to generate much larger data sets. "This is just the tip of the iceberg," he says of the published work. "Within ten years I'm convinced we'll be imaging the activity of thousands of neurons in a living brain and tracing tens of thousands of connections between them."

MORE INFO:
www.psc.edu/science/2011/nanomappers



PUTTING GENES TOGETHER REALLY FAST



PSC's newest supercomputer, Blacklight, is helping to break open a potential bottleneck in processing and analysis of DNA sequence data

It didn't take long for Blacklight to show its mettle as a tool for genome sequencing. PSC's newest supercomputer, a resource of XSEDE (see p. 5), came online as a production system in October 2010, and due to its availability, two projects involving genomics — a science that has in the last few years shifted into data-intensive overdrive — made remarkable progress.

New sequencing instruments, hardware technologies that "read" sequences of DNA and decipher the order of nucleotide bases — A, G, C and T (adenine, guanine, cytosine and thymine) — have begun to produce data at unprecedented speed. "Within the last three to five years," says Cecilia Lo, chair of the University of Pittsburgh School of Medicine's Department of Developmental Biology, "new sequencers have come on line, carrying out sequencing that is referred to as 'next-generation sequencing.' What used to take years with capillary sequencing can now be accomplished in a matter of one or two weeks."

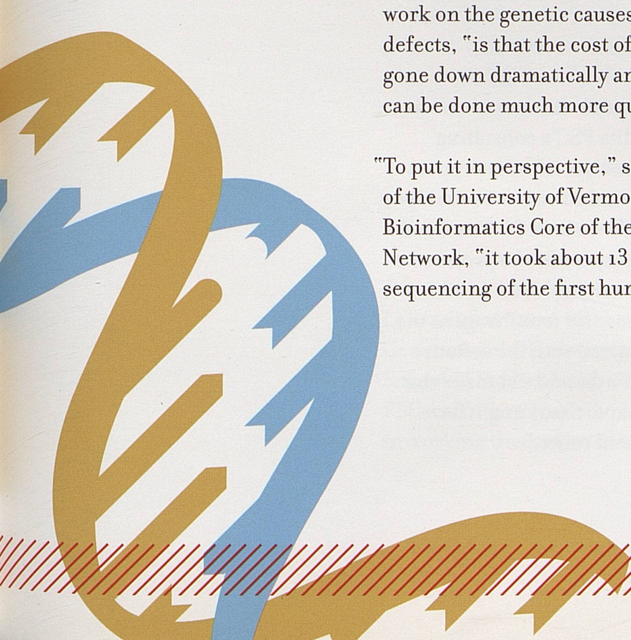
The essential difference is long versus short reads. Previous sequencers did reads of about 300 to 500 and sometimes up to 1000 bases. The new technologies do reads of 50 to 100 bases. "The result," says Lo, who has used Blacklight for her work on the genetic causes of congenital heart defects, "is that the cost of sequencing per base has gone down dramatically and the sequencing runs can be done much more quickly."

"To put it in perspective," says James Vincent of the University of Vermont, who directs the Bioinformatics Core of the Vermont Genetics Network, "it took about 13 years to complete the sequencing of the first human genome. The new instruments can sequence two human genomes in a single run." As part of a team of bioinformatics scientists collaborating through the Northeast Cyberinfrastructure Consortium (NECC), Vincent is using Blacklight to assemble the genome of the little skate (*Leucoraja erinacea*), a fish species of the northwestern Atlantic. "The amount of sequencing data that can be generated from a single instrument," he adds, "has for several years been doubling every four or five months."

While these skyrocketing quantities of sequence data are a blessing for biological science, they pose the problem and challenge of a potentially stifling analytical bottleneck. Once a sequencing instrument has produced millions or, as the case may be, billions of reads from an organism's DNA, researchers face the task of assembling them into a complete genome. Blacklight is helping to break this bottleneck.

For Lo's work, involving mouse genome data, her collaborators used Blacklight to process over 700 million reads and assemble them into a whole genome in eight hours. This compares to about two weeks on a laboratory-based cluster system she and her collaborators had been using.

For Vincent, the Blacklight advantage is perhaps even greater. With NECC's little skate project, he worked for several months on other computing systems before coming to PSC. "Within a week," says Vincent, "90-percent of my problems were solved." With billions of 100-base reads, he was able to complete a *de novo* assembly of the little skate genome in weeks, a large step toward a complete analyzed genome, progress that had eluded him on other systems for nearly a year.



THE RESEARCHERS

James Vincent (left)
UNIVERSITY OF VERMONT

Phil Blood
XSEDE ADVANCED USER SUPPORT CONSULTANT, PSC

THE LITTLE SKATE

As a sequencing project, the little skate isn't just any fish that hadn't yet had its genome sequenced. It's one of only 11 non-mammals selected by the NIH as a "model organism," organisms that have "the greatest potential to fill crucial gaps in human biomedical knowledge." Model organisms are often used as a reference for better understanding human disease conditions. The skate, for instance, shares characteristics with the human immune, circulatory and nervous systems.

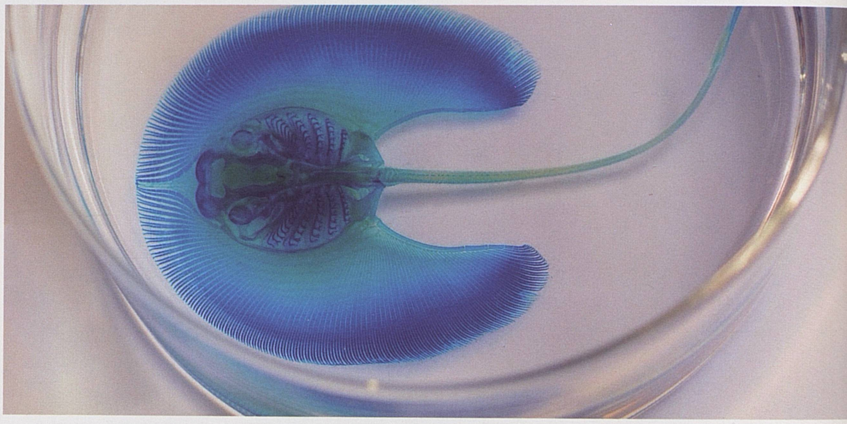
With an American Recovery and Reinvestment Act grant in 2009, NECC was formed to create a high-speed fiber-optic network in five northeast states. It also links five institutions with bioinformatics research programs — Mount Desert Island Biological Laboratory (MDIBL) in Maine, the University of Delaware, Dartmouth College in New Hampshire, the University of Rhode Island and the University of Vermont. When NECC was established, says Vincent, the plan to sequence the entire genome of the little skate, originally a project at MDIBL, gained momentum: "This is the kind of project we anticipated when building our network. It's both an excellent demonstration project for the use of the infrastructure, and at the same time, it's a superb scientific project."

A large part of the challenge is the lack of a reference genome. "It's called *de novo* assembly," says Vincent. "We have to create the little skate genome from scratch. This particular branch down the tree of life fits a niche that doesn't yet have sequenced genomes."

The skate genome is 3.4 billion bases, a little larger than the human genome, and the task was to take billions of 100-base reads — in which many of the bases, sometimes as many as 99, overlap with the next read — and match them with each other in the right order. "The sequencing instrument gives you billions of tiny pieces of a long continuous DNA string," says Vincent, "and to create a draft genome *de novo*, you have to put those little pieces back together."

Using software called ABySS, Vincent brought the project to PSC for NECC in 2011. ABySS's algorithm for genome assembly, he explains, builds a graph of the relationships from all the reads in memory. Some parts of the job require only one or a small number of processors, while others exploit massive parallelism — many processors at once. Because of this, says Vincent, it's often necessary to move back and forth between a massively parallel cluster and a single-processor machine with very large memory. "Blacklight's shared memory makes all this go away."

"Memory is shared across all the nodes, so you can treat it like a traditional cluster, or you can access all the memory you have allocated from a single node, which acts like a single, large-memory machine. You need both of these to complete the ABySS job, and you can do that all at once on Blacklight. This machine made running ABySS easy."



Little skate in a lab dish
Credit: Mount Desert Island Biological Laboratory

As a result, Vincent was able to complete a draft genome of the little skate, which he and his collaborators are now analyzing in comparison with another little skate draft genome — done by Ben King at MDIBL with different software.

As much as Blacklight's shared-memory architecture, says Vincent, PSC staff made the project go.

As much as Blacklight's shared-memory architecture, Vincent credits PSC's consulting staff, in particular XSEDE advanced user support consultant Phil Blood. "I wouldn't have been able to do anything on Blacklight without PSC staff and Phil Blood in particular. They made the project go. Phil took a real interest and solved a lot of things that were hard for me. He found bugs in the software and got them resolved with the software authors. I'd worked for months and not made that progress. Without Phil's expertise, I might have given up and gone a different route."

TRACKING THE GENES FOR DEFECTIVE HEARTS

Cecilia Lo and her colleagues would like to identify the genes involved in human congenital heart disease. Using genetically modified mice, her research group aims to find gene mutations that cause problems in cardiac development that can lead to structural heart defects — such as holes in the walls of the heart or abnormal connection of the aorta or pulmonary artery — defects that affect almost one-percent of live births and can cause newborn infant death.

Ultimately, the goal is a "diagnostic chip" for human congenital heart disease. "This is the age of personalized medicine," says Lo, founding chair of her department, one of a handful of developmental biology departments nationwide, "and that's where medicine is headed. Such a chip would provide the possibility to retrieve sequence information on many if not all of the genes involved in structural heart disease."

In the future, each person with congenital heart disease coming to the hospital for treatment, she explains, would have their blood drawn to obtain DNA. Sequencing analysis with this chip could determine if the patient has cardiac-related gene mutations, and a treatment plan could be customized to the patients genetic make up. "We want to understand how these genes contribute to structural heart disease, influence disease progression, and affect long-term outcome. Whether medication or a surgical approach, you would be able to optimize and personalize the medical care of each patient based on the individual's specific genotype."

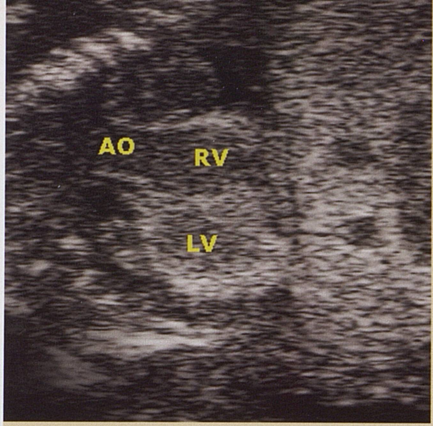
THE RESEARCHERS

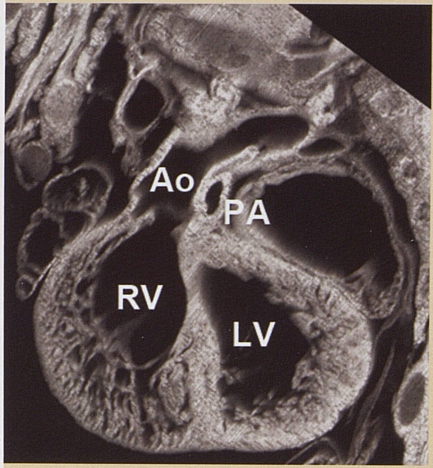
Cecilia Lo
UNIVERSITY OF PITTSBURGH SCHOOL OF MEDICINE (top)

Michael Barmada
UNIVERSITY OF PITTSBURGH GRADUATE SCHOOL OF PUBLIC HEALTH

To find the core set of heart-defect related genes, Lo and colleagues are screening over 100,000 mutant mouse fetuses over five years. When they see a heart defect, using noninvasive *in utero* ultrasound imaging, they follow up by sequencing the genome of that mouse — for comparison to the reference genome of a normal, healthy mouse. "In this way," says Lo, "we should be able to identify the mutations involved in the heart disease."

A next-generation sequencer can rapidly provide sequence data for the entire genome of each mutant mouse. To assemble the reads into a complete genome, one approach is to break the data into chunks, says Michael Barmada, one of Lo's collaborators, from the Department of Human Genetics at the University of Pittsburgh's Graduate School of Public Health. "We worked closely with PSC staff," he says, "who have been really helpful, in working this out on Blacklight."





CONGENITAL HEART DEFECTS
Ultrasound imaging (top) of a genetically-modified mouse *in utero* showed the aorta (AO) emerging from the right ventricle (RV). Subsequent microscopic histopathology (bottom) showed a double-outlet right ventricle with pulmonary atresia — the aorta and a very small pulmonary artery (PA) emerge from the right ventricle.

After testing and benchmarking several approaches, Barmada split the sequence data into several independent chunks that ran concurrently on 1000 Blacklight cores, with the result that assembly of the genome for one mutant mouse took only eight hours. "This was taking at least a week-and-a-half," says Barmada, "on a 24-core machine in our lab."

"Blacklight is allowing us to do things that would be very difficult to do otherwise."

"Given the many mouse mutants awaiting analysis, we have a huge amount of sequencing data," says Lo, "that will need to be mapped back to the mouse reference genome. Blacklight is allowing us to do things that would be very difficult to do otherwise."

MORE INFO:
www.psc.edu/science/2011/sequencing

MINING THE WORD HOARD

Photo credit: Greg Hanneman

THE RESEARCHER

Noah Smith CARNEGIE MELLON UNIVERSITY

With Blacklight's shared memory, Carnegie Mellon scientists are upping the ante of what's possible with natural language processing

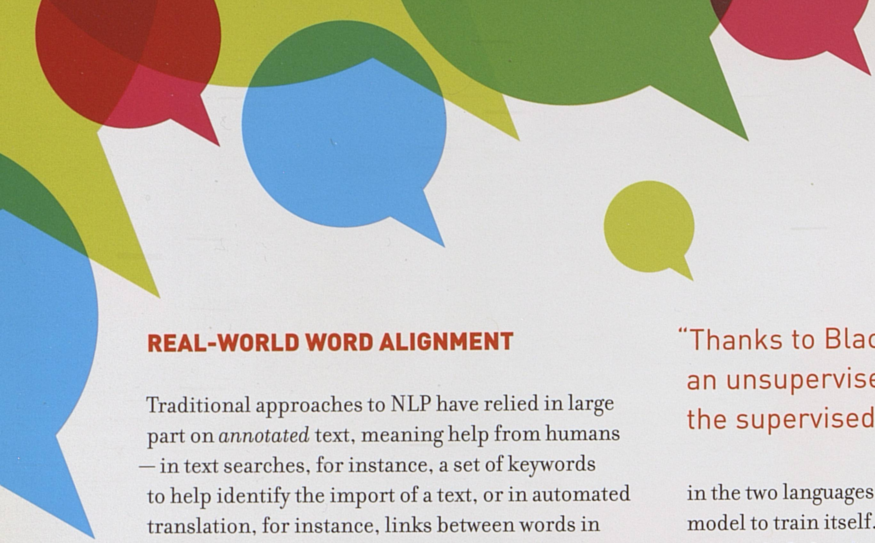
For most of us words are how we communicate and, mostly, we don't give much thought to them beyond that. Sometimes they occur to us spontaneously, in delight or sorrow. Other times we use them in sentences carefully crafted to express nuances of thought. For computer scientists in the field of natural language processing (NLP), however, words are also data, and there's plenty to go around.

The World Wide Web has become an expanding, limitless repository of text, billions and billions of words, and for Noah Smith and his colleagues it's a treasure trove — to sift through, ask questions, test better ways to translate languages, and sometimes to make forecasts about collective human behavior. "The general area we work in is natural language processing," says Smith, associate professor in the Language Technologies Institute at Carnegie Mellon University, one of the world's leading centers in using computers to solve language-related problems.

"You can imagine anything from more intelligent search engines to answer your questions," says Smith, "to systems that translate automatically from one language to another." In late 2010, while Blacklight, PSC's newest supercomputer (see p. 4), was undergoing shakedown, Smith and his colleagues were experimenting with the new system, work that bore fruit — four papers within six months, in diverse areas of NLP.

"Blacklight has been a very useful resource for us," says Smith. "We can incorporate deeper ideas about how language works, and we can estimate these more complex models on more data." Blacklight's shared memory has been crucial, he observes, because his large-scale models use iterative algorithms that look at the same data over and over again. "Shared memory lets us use many processors in parallel without having to worry about the overhead of passing data over the network or moving model information around."

A recurring theme of Smith's modeling, and one of the reasons Blacklight's shared memory has opened doors in his work, is an "unsupervised" approach to text data, as exemplified by his group's recent work on word alignment, an important component of automated language translation.



REAL-WORLD WORD ALIGNMENT

Traditional approaches to NLP have relied in large part on *annotated* text, meaning help from humans — in text searches, for instance, a set of keywords to help identify the import of a text, or in automated translation, for instance, links between words in English sentences and their Chinese translations. In general, Smith's work contrasts with this. "Our approach," he says, "is to discover the structures of interest from large amounts of *unannotated* text data."

"Unsupervised" is a general term for this approach, which allows the model to start from scratch and sift through real-world text, without expensive expert annotations, to build connections in the data for the task it undertakes to accomplish. This has the advantage of not limiting tasks based on whether annotated text is available and, further, offers the potential to uncover linguistic connections less biased by previous thinking. With translation in particular, observes Smith, unsupervised approaches have become more feasible as huge amounts of text have accumulated on the web, with the United Nations as a prime example.

"Everything that happens at the UN has to be translated by expert translators into the major languages that people speak around the world. The result is what we call 'parallel documents' — text in English and corresponding text in Chinese, Arabic and other languages. This data is freely available to everybody."

For Smith, these parallel documents — and the availability of Blacklight — made it feasible to try an experiment in word alignment, the part of translation in which a model builds statistically-based maps of connections between words in two languages. Smith's project built alignments between English and Czech, English and Chinese and English and Ordu — languages very different from each other. Czech, for instance, notes Smith, has complex morphologies in which the verb changes depending on whether the subject of the sentence is masculine or feminine, or singular or plural.

"The more data you have," says Smith, "the better you can do with word alignment, but likewise it becomes more and more expensive computationally." In some recent work, for this reason, word alignment has relied on human experts to draw links between a subset of words in the two languages as a starting point for the model to train itself. "It gives a nice clean gold standard of what the alignments look like. The problem, of course, is human intervention is costly and you can do it only for a small amount of data."

"Thanks to Blacklight we were able to train an unsupervised model that outperforms the supervised approaches."

With Blacklight's ability to hold large amounts of data in shared memory, Smith's unsupervised word-alignment model in all three test cases outperformed other unsupervised alignment models. By a variety of measures, furthermore, including using the model in automated translation programs, his model outperformed supervised approaches, which hadn't been accomplished in previous natural language modeling.

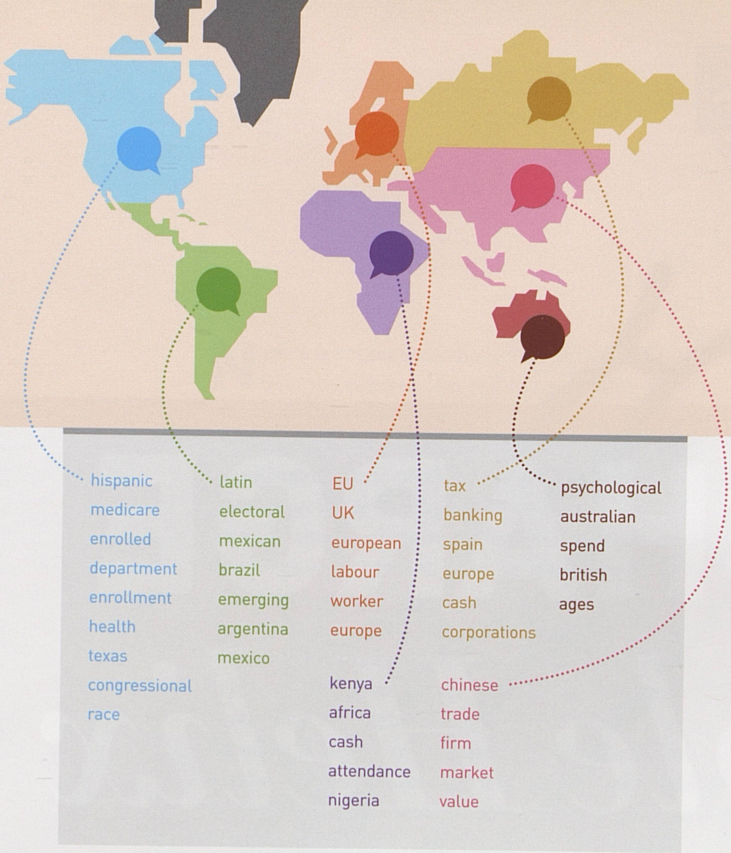
"Thanks to Blacklight," says Smith, "we were able to train an unsupervised model that outperforms the supervised approaches. This is because we were using massive amounts of data, along with some sophisticated statistical modeling techniques that had been applied before only in supervised cases. It was an obvious gap, and it was because the amount of computing was prohibitive that people hadn't tried this before."

FORECASTING SCHOLARLY IMPACT

How much can the word-data of many texts make it possible to predict how people will respond to other similar texts? For a few years, Smith and his colleague Bryan Routledge from CMU's Tepper School of Business have explored "text-driven forecasting" — the ability of statistical modeling to discern features in text that can reliably forecast human responses.

With this relatively new application of NLP, they have, for instance, shown that the prevalence of words of identifiable characteristics and flavors in movie reviews can predict with statistical validity whether the movie will make a profit on opening weekend. Another of their projects found that language features of corporate financial reports can predict the volatility of that corporation's stock price over the next year.

With Blacklight, Smith and Routledge and their collaborators have been testing a more ambitious possibility: Can you forecast the scholarly impact



SCHOLARLY IMPACT BY WORLD REGION
A part of the Smith group's study with the National Bureau of Economic Research database of papers mapped downloads to world regions and tracked correlations between text content and download "popularity" by region. This chart shows the most informative single-word features for each of seven world regions.

of scientific articles — with "impact" measured as how often an article is cited in other papers or downloaded from the web — from its text content?

For this project, the researchers used two large datasets of scientific papers. The National Bureau of Economic Research (NBER) provided download data on papers, from approximately 1,000 economists, posted in the NBER online archive. A second dataset comprised papers and related citation data from the Association for Computational Linguistics (ACL).

"This gives us paired data," says Smith. "We have documents, the research papers, and a response that came in later. How much did people download a paper, or cite it? It's a way to measure a response within the community."

Unlike translation, this is a supervised problem because the model learns from what happened in the past. For the NBER data, for instance, the model trained itself on 10 years worth of papers, 1999 to 2009, to learn relationships between text content and the number of downloads. The researchers then tested the model's ability, based on what it learned from the historical data, to predict download response for papers held out from the training database.

Compared to other data associated with scientific papers — such as author's name, the category of topic, what journal, the text-content based predictions were significantly more accurate. "The accuracy went up when we used these newer techniques," says Smith. "Nobody had framed this problem quite this way before."

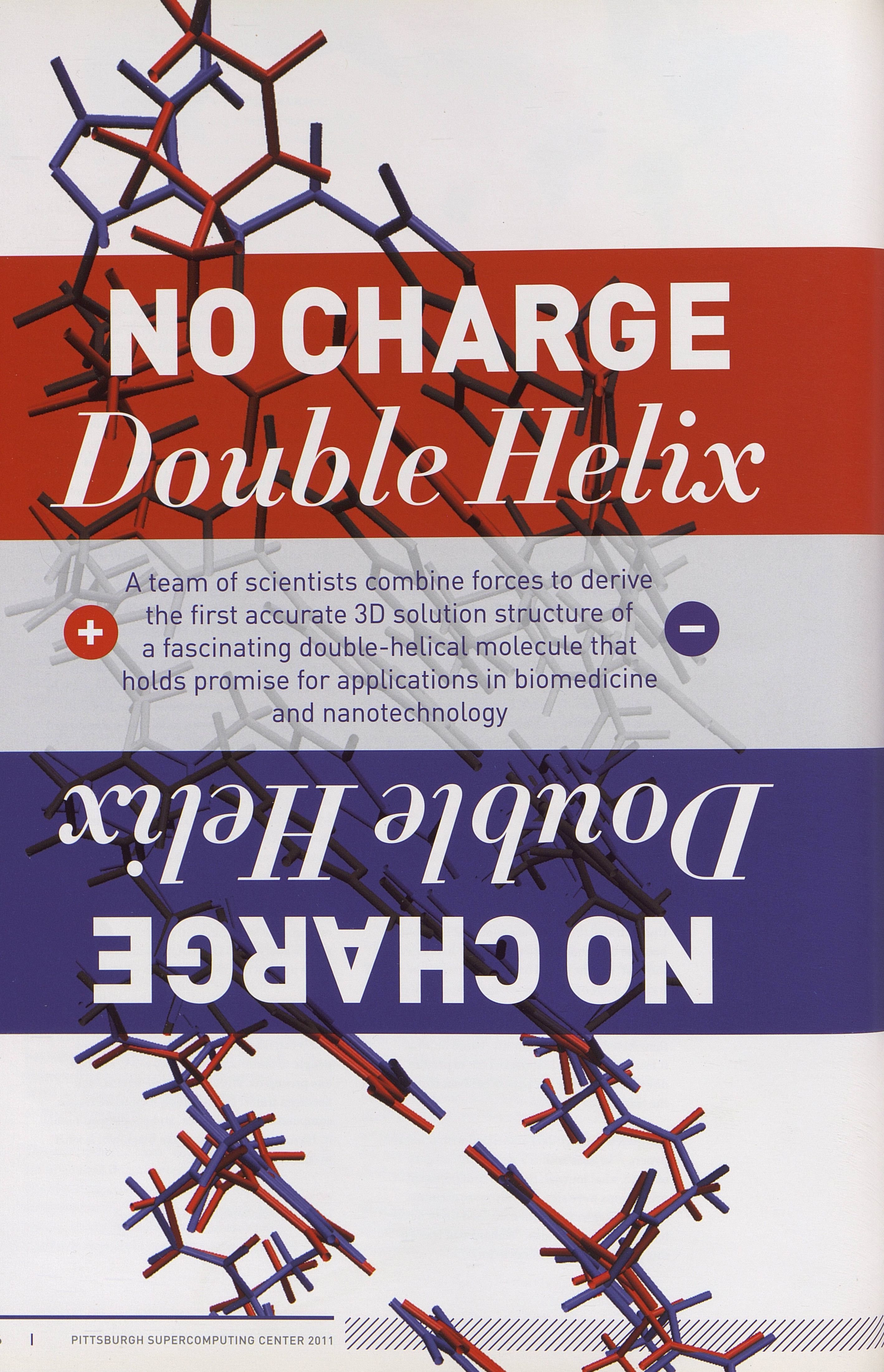
Despite its not being an unsupervised model, this project was computationally demanding, says Smith, because of its high number of dimensions. "We're looking at a very large set of clues from the input text about the future." And the model is "discriminative" — designed to learn to do a specific task and get it right: "The estimation procedure is computationally expensive, and we've run a number of experiments, in which we've divided the data in different ways, to see what ideas work."

Beyond the impact predictions, this model also tracks how scholarly impact of a paper may change over time, which Smith calls "time series changes." He compares this dimension of the model to related approaches that track "term frequency" in texts, a count of how often a word appears in a given time period. Term frequency by itself, Smith believes, is less revealing than a measurement, such as downloads or citations, that connects term frequency to impact. "While the time series is much more computationally expensive than counting words, it gives a fuller understanding of the community response."

The implications of this kind of text-driven forecasting extend to such possibilities as helping busy people make intelligent choices of what to read. "Can we learn to take this messy and unstructured data and with computers turn it into some deeper meaning representation?" asks Smith. "A human couldn't read all the papers that come out in a year on the NBER website. They can't read fast enough. But with this kind of model we can look and do analysis on that large amount of data and come up with trends that tell you what's going on."

"These estimation procedures," he continues, "become expressed in very complicated algorithms, and it can take multiple graduate courses for people to understand how they work, even on one processor. With real data scenarios, this becomes almost frighteningly expensive from a computational standpoint, and people won't touch it. Having an architecture like Blacklight is what makes this work possible."

MORE INFO:
www.psc.edu/science/2011/language



NO CHARGE *Double Helix*

A team of scientists combine forces to derive the first accurate 3D solution structure of a fascinating double-helical molecule that holds promise for applications in biomedicine and nanotechnology

NO CHARGE *Double Helix*

Nature is sophisticated, says Catalina Achim. Over the course of several billion years, it has evolved remarkably efficient ways to transfer electrons from atom to atom within living organisms to produce energy from food. Our energy for getting up in the morning, for work and for pleasure all comes from these processes of "controlled burning" that depend on electron transfer.

"Just like we transfer electricity through power lines to heat and light our homes, we transfer electrons in our bodies to metabolize food," says Achim, associate professor of chemistry at Carnegie Mellon University. "But nature does it very efficiently while we don't yet know how to take oil and make our cars go without wasting a lot of energy. We know some of the basics of how electron transfer works, and many scientists study these processes so we can learn to apply them."

To that end, Achim has worked with a team and turned to XSEDE resources, including PSC scientist Marcela Madrid, and Pople, PSC's SGI Altix system, to solve the structure of a fascinating "bio-mimetic" molecule called PNA, *peptide nucleic acid*. Their aim is to use PNA as a molecular "scaffold" — to organize metal ions so that they can transfer electrons as efficiently as biological electron-transfer molecules. Although PNA doesn't exist in nature, it's a close cousin to DNA and for Achim's purposes has a special advantage — it doesn't have charge.

DNA's helical strands have negative charges, due to repeating phosphate groups within the backbone, and when the four A-G-C-T bases pair up to form the DNA double-stranded helix, there's electrostatic repulsion between the strands. "The DNA structure is overall always negatively charged," says Achim.

DNA interested Achim as a scaffold for metal ions, but the negative charge and its sensitivity to chemical degradation in living organisms made it unsuitable. To circumvent these limitations, Achim and her colleagues turned to PNA, which they synthesize in her Carnegie Mellon laboratory — by replacing the DNA backbone with peptide-like groups. The resulting neutral double helix is more stable and better suited for research in electron transfer.

The necessary first step for electron-transfer research, though, was an accurate 3D structure of PNA in solution with water. For this task, Achim and her collaborators teamed with Madrid, an expert in molecular dynamics (MD), an approach that can find a molecule's structure by computing the forces between its atoms. The results of their work, reported in *Molecular Biosystems* (2010), a journal of the Royal Society of Chemistry, provided for the first time the PNA structure in solution.

From this starting place, Achim and her colleagues can investigate potential applications of PNA that can open new understanding of electron transfer. Achim foresees, for instance, that PNA could be used to create "nanowires" — 100,000 times finer than a human hair — for quantum circuitry, in which the quantum characteristics of electrons hopping and tunneling can lead to electronics much faster than today's integrated circuitry.

THE RESEARCHERS

[top]

Catalina Achim

(third from left) with her laboratory group

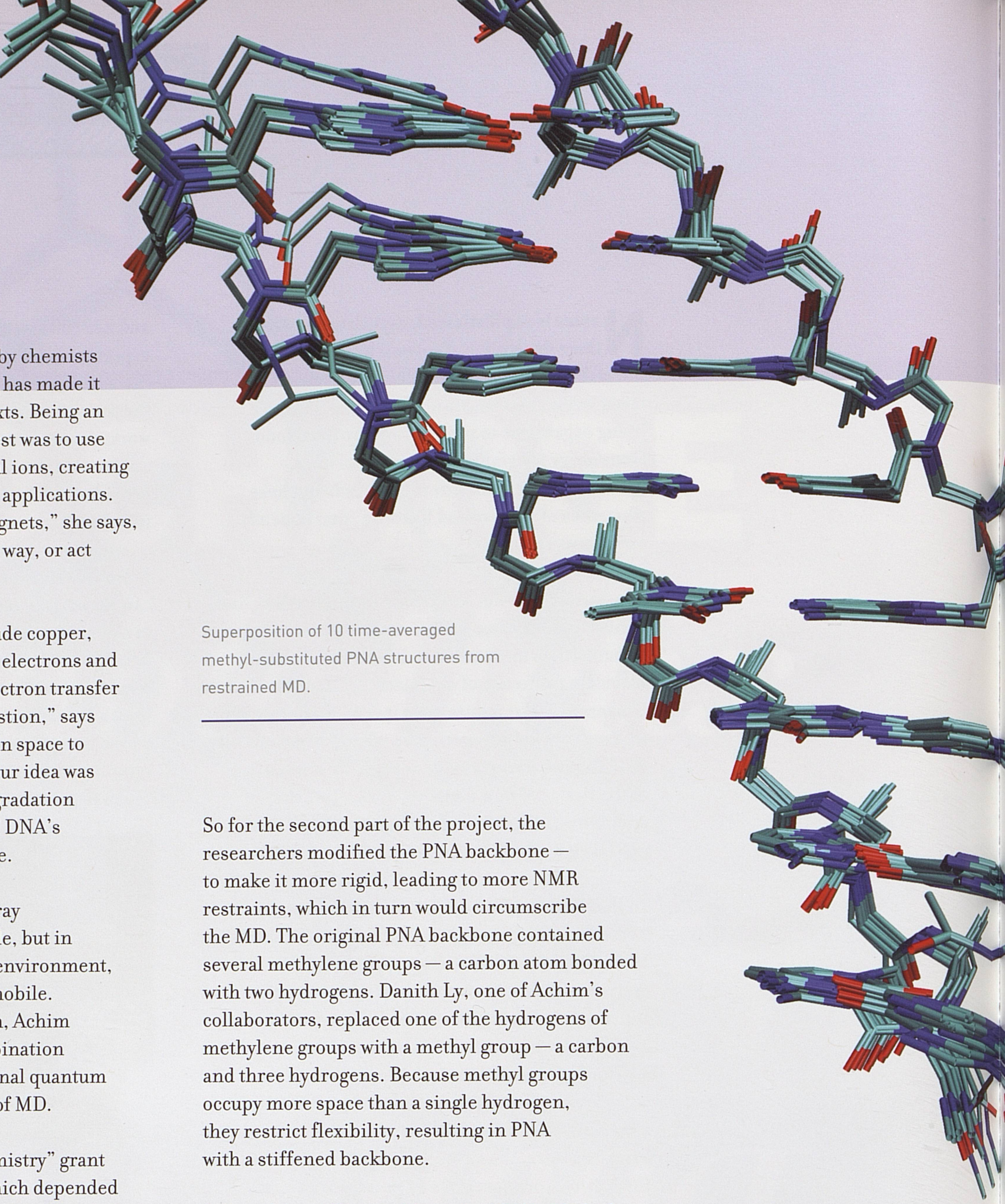
CARNEGIE MELLON UNIVERSITY

[bottom]

Marcela Madrid

PITTSBURGH SUPERCOMPUTING CENTER, XSEDE ADVANCED USER SUPPORT SERVICES

"Our collaboration with PSC was very beneficial," says Achim, "not only for the research itself but also for educational purposes. Working with Marcela Madrid, my graduate students learned how to do molecular dynamics simulations."



Superposition of 10 time-averaged methyl-substituted PNA structures from restrained MD.

A STIFFENED BACKBONE

PNA was first synthesized in 1991 by chemists in Denmark. Its similarity to DNA has made it a subject of study in various contexts. Being an inorganic chemist, Achim's interest was to use it as a platform for transition metal ions, creating molecular complexes with various applications. "They may behave as molecular magnets," she says, "or transport charge in a particular way, or act as catalysts."

The transition metals, which include copper, iron or ruthenium, easily transfer electrons and hence are ideal for the study of electron transfer via a molecular scaffold. "Our question," says Achim, "was 'What's the best way in space to organize transition metal ions?' Our idea was 'Let's use a scaffold that resists degradation and is easily modified in the lab.'" DNA's drawbacks brought PNA to the fore.

A static structure of PNA, from x-ray crystallographic study was available, but in biological applications in a water environment, PNA isn't static; it's flexible and mobile. To find its 3D structure in solution, Achim and her collaborators used a combination of NMR spectroscopy, computational quantum chemistry and a specialized form of MD.

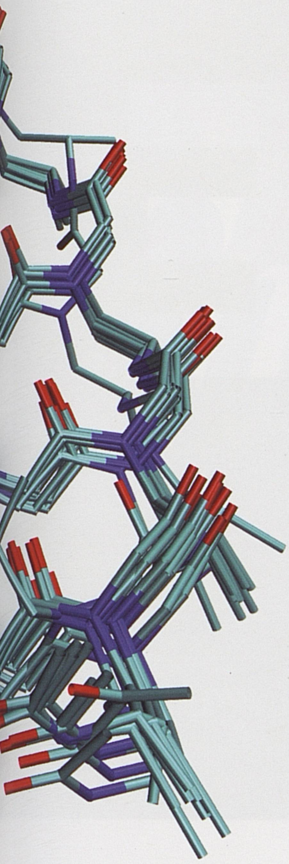
A "Collaborative Research in Chemistry" grant from NSF supported this work, which depended on teamwork among Achim, Madrid and Achim's partners at Carnegie Mellon and Duke University. First, Achim and her students synthesized several versions of PNA, each with a different chemical structure and flexibility in solution. A graduate student in Achim's lab worked with Carnegie Mellon chemist Roberto Gil on the NMR spectroscopy of the synthesized PNAs, which provided a matrix of distances between protons in the molecules.

With software called AMBER, Madrid did "restrained" MD, a technique by which it is possible to find a family of 3D structures that fit the NMR data. "We did about ten of these simulations," says Madrid, "and got a family of structures that are compatible with the experimental data." This initial family of structures, however, had backbones that varied significantly from each other. These variations, the researchers realized, were attributable to intrinsic flexibility of the PNA backbone, which in turn limits the number of restraints that NMR can determine.

So for the second part of the project, the researchers modified the PNA backbone — to make it more rigid, leading to more NMR restraints, which in turn would circumscribe the MD. The original PNA backbone contained several methylene groups — a carbon atom bonded with two hydrogens. Danith Ly, one of Achim's collaborators, replaced one of the hydrogens of methylene groups with a methyl group — a carbon and three hydrogens. Because methyl groups occupy more space than a single hydrogen, they restrict flexibility, resulting in PNA with a stiffened backbone.

Before doing MD with this revised PNA, Madrid calculated the changed charges between atoms due to the methyl group, which she needed as input to the MD "force fields" — mathematical expressions that facilitate MD by representing the quantum-mechanical energies between atoms. For this charge calculation, working with Achim's grad students, Madrid turned to GAUSSIAN98 — a quantum chemistry program originally developed by Carnegie Mellon theoretical chemist John Pople, for which he received the 1998 Nobel Prize in Chemistry.

With revised charges and new NMR restraints, another round of MD computations produced a family of 3D structures that fit the NMR data and also fit well with the available crystallographic structure with much less variation in the backbone. "Handling the limitations imposed by the backbone," says Achim, "was frustrating at the beginning, but it led us to the substitution with the methyl group and to understanding the PNA structure in very much detail. That's the way research goes; you don't know



Fast turnaround of the many computations was essential to the success of the project.

the answers when you start but everything falls in place like the letters in an interesting acrostic."

The fast turnaround of the many computations, possible with the supercomputer named in honor of Pople and the help of Madrid, says Achim, was essential to the success of the project. "We are always anxious to get results and push the limits of our knowledge. Access to PSC computer resources and, even more, the help of Marcela made a huge difference in how fast we got the answers we were seeking."

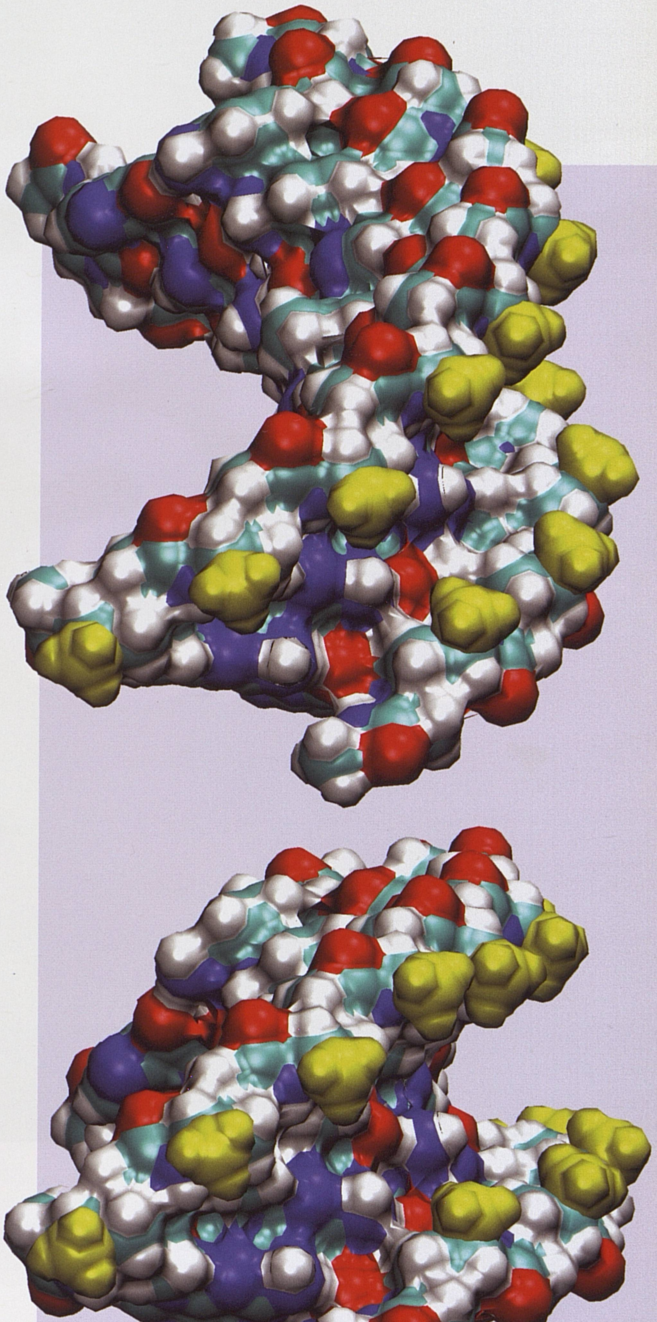
LOOKING AHEAD

With the structure of their modified PNA in hand, Achim and her colleagues are proceeding with studies of metal-containing PNAs and electron transfer. "Now we can think about how the fundamental knowledge we have acquired can be translated into possible applications, both in nanotechnology as well in biology."

In recent work, Achim and her students have created PNA molecules with copper, nickel and iron. David Waldeck, a University of Pittsburgh chemist, deposits those molecules on a gold electrode, creating self-assembled PNA monolayers. Waldeck then studies the electron flow through them. "He found that the mechanism of charge transfer in PNA," says Achim, "is tunneling at short distances and hopping at long distances."

Achim foresees that PSC's new shared-memory system, Blacklight, will further advance her work with bio-mimetic molecules. "Based on the structures we have solved together, we have ideas about how to design additional nucleic-acid based structures. Those would be more complex, and require more computational effort and larger resources, so it is very exciting that these things come together. I look forward to continuing the collaboration, and to using this new resource."

MORE INFO:
www.psc.edu/science/2011/helix



Side (top) and axial (bottom) view of PNA simulated by Achim, Madrid and colleagues. The substituted methyl groups are yellow. Other atoms are hydrogen (white), oxygen (red), nitrogen (blue), carbon (teal).

SUPER MASSIVE

With MassiveBlack, the largest cosmological simulation of its kind to date, and a new approach to visualizing the results, enabled by PSC's Blacklight, astrophysicists solved a puzzle about how some of the first black holes in the universe became supermassive in such a short time

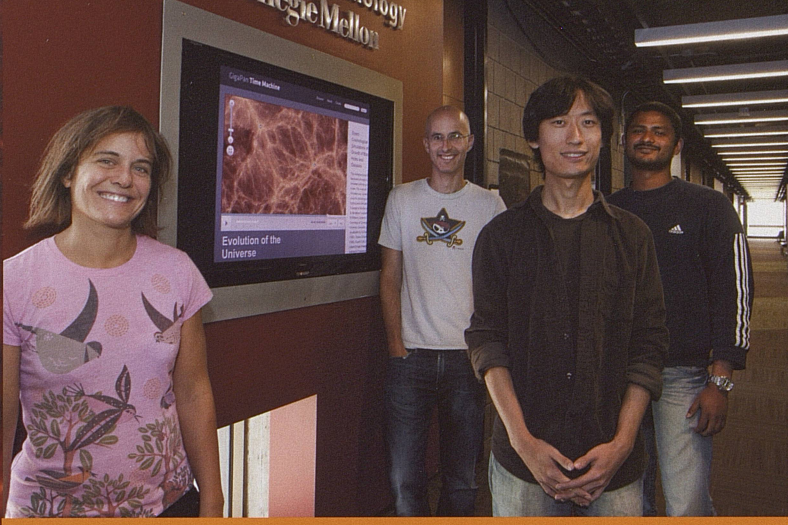
"How did we get these huge monsters so early on?" asks Tiziana Di Matteo. The Carnegie Mellon astrophysicist is referring to supermassive black holes, which — astrophysicists now know — reside at the center of every large galaxy. These cosmic behemoths, with mass that can be many billion times that of the sun, swallow huge quantities of gas and form the gravitational cores that have structured matter into galaxies. Although themselves invisible, they signal their presence by the quasars they spawn, as inward-drawn gas heats and radiates light that can be a hundred-trillion times brighter than the sun.

Di Matteo's question responds to recent astronomical observations, such as the Sloan Digital Sky Survey, that have discovered quasars associated with supermassive black holes in the first billion years after the big bang. The existence of black holes *per se* isn't surprising, but supermassive ones at these infant stages of the universe, currently 13.6 billion years old, present a challenge for the reigning "cold dark matter" model of how the universe evolved. "If you write the equations for how galaxies and black holes form," says Rupert Croft, Di Matteo's Carnegie Mellon colleague, "it doesn't seem possible that these huge masses could form that early."

To resolve this puzzle, Di Matteo, Croft and their collaborators turned to supercomputing. "Even before the recent quasar observations," says Di Matteo, "it's been a pressing question. When and how did the first black holes form?" To get some answers, Di Matteo and Croft and the rest of their group in 2010 mounted a very large-scale simulation with Kraken, a Cray XT5 XSEDE (see p. 5) resource at the University of Tennessee (NICS). They called their simulation MassiveBlack.

THE RESEARCHERS

[from left to right]
Tiziana Di Matteo
Rupert Croft
Yu Feng
Nishikanta Khandai



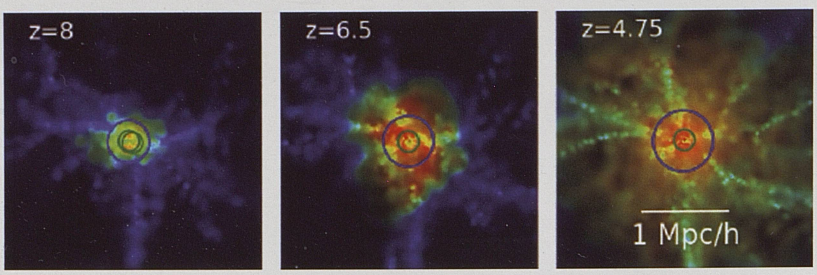
Growth Spurt

Using all of Kraken's compute cores, nearly 100,000, with post-doctoral fellow Nishikanta Khandai doing most of the hands-on computing, the researchers simulated a huge volume of space (a cube of .75 gigaparsecs per side, about 2.5 billion light years). Within this volume, MassiveBlack included 65.5 billion particles to represent matter, as the universe evolved from 10 million years after the big bang through the period of early structure formation and emergence of the first galaxies and quasars to when it was 1.3 billion years old.

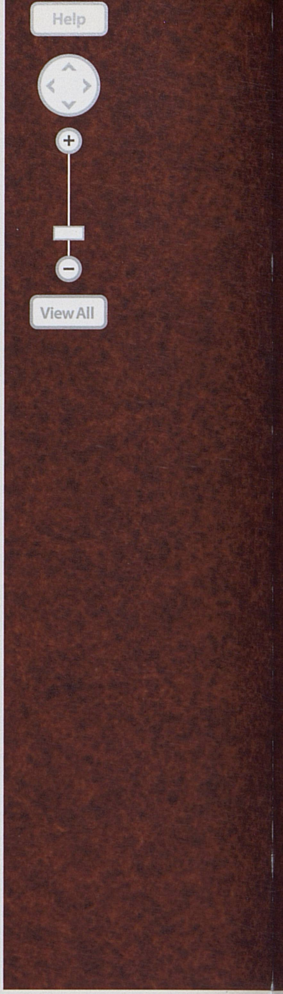
The largest cosmological simulation of its kind (smooth particle hydrodynamic) to date, MassiveBlack produced a massive amount of data, for which the researchers turned to Blacklight at PSC, the world's largest shared-memory computing system. Blacklight made it possible to hold a snapshot of the entire simulation volume, three terabytes of data, in memory at one time.

With help from this new approach to visualization, the researchers identified a physical phenomenon that goes far toward explaining the existence of supermassive black holes so early in the universe.

"We were able to show," says Di Matteo, "that in regions of high density the gas comes straight into the center of the black hole, extremely fast, and in these places we see that the black holes grow really, really quickly." The researchers call this phenomenon "cold gas flows." It had been seen in other simulations and had been gaining acceptance as a phenomenon involved in galaxy formation, but only at much lower redshifts, when the universe is older, not within the first billion years. "This was the first simulation," adds Di Matteo, "to see this at high redshift."



FAST FOOD FOR THE MONSTER Three snapshots from MassiveBlack at three different redshifts (higher redshift represents earlier time) show evolution of a quasar associated with a supermassive black hole within the first billion years of the universe. Gas distribution is color coded by temperature (blue through red). Cold streams of gas (green) penetrate the dark matter "halo" (blue circle) of the black hole (green circle) at the galaxy center.



EVOLUTION OF A SIMULATION

To run MassiveBlack, Di Matteo, Croft and colleagues used a cosmological simulation code (called P-GADGET) with proven ability to track the physics of black holes and quasars along with galaxy formation as the universe evolves. Di Matteo led a simulation five years earlier, a large run with GADGET on PSC's then largest system, BigBen, that was the first simulation to include the physics of black holes and to run at sufficiently fine resolution to track their formation.

Using 2,000 of BigBen's Cray XT3 processors in parallel over four weeks of run time, that simulation tracked a volume of cosmos (a cube of 33 megaparsecs) large for the time, but more than 1,000 times smaller than MassiveBlack. The findings — on the relationships between black holes and galaxy structure and the feedback process by which black holes eventually shut off their associated quasars — offered many new insights. "We're still publishing papers from that simulation," says Di Matteo. "It was very rich in science."

MassiveBlack, however, metaphorically speaking, was another universe entirely, requiring much more extensive computing, a gigantic run. The starting requirement was a much larger volume of space, since supermassive black holes are very rare objects in the early universe. "If you looked at that volume in the present day," says Croft, "there would be about a million Milky Way size galaxies. But in this early epoch, there would be only a handful of quasars with black holes of a billion solar masses."

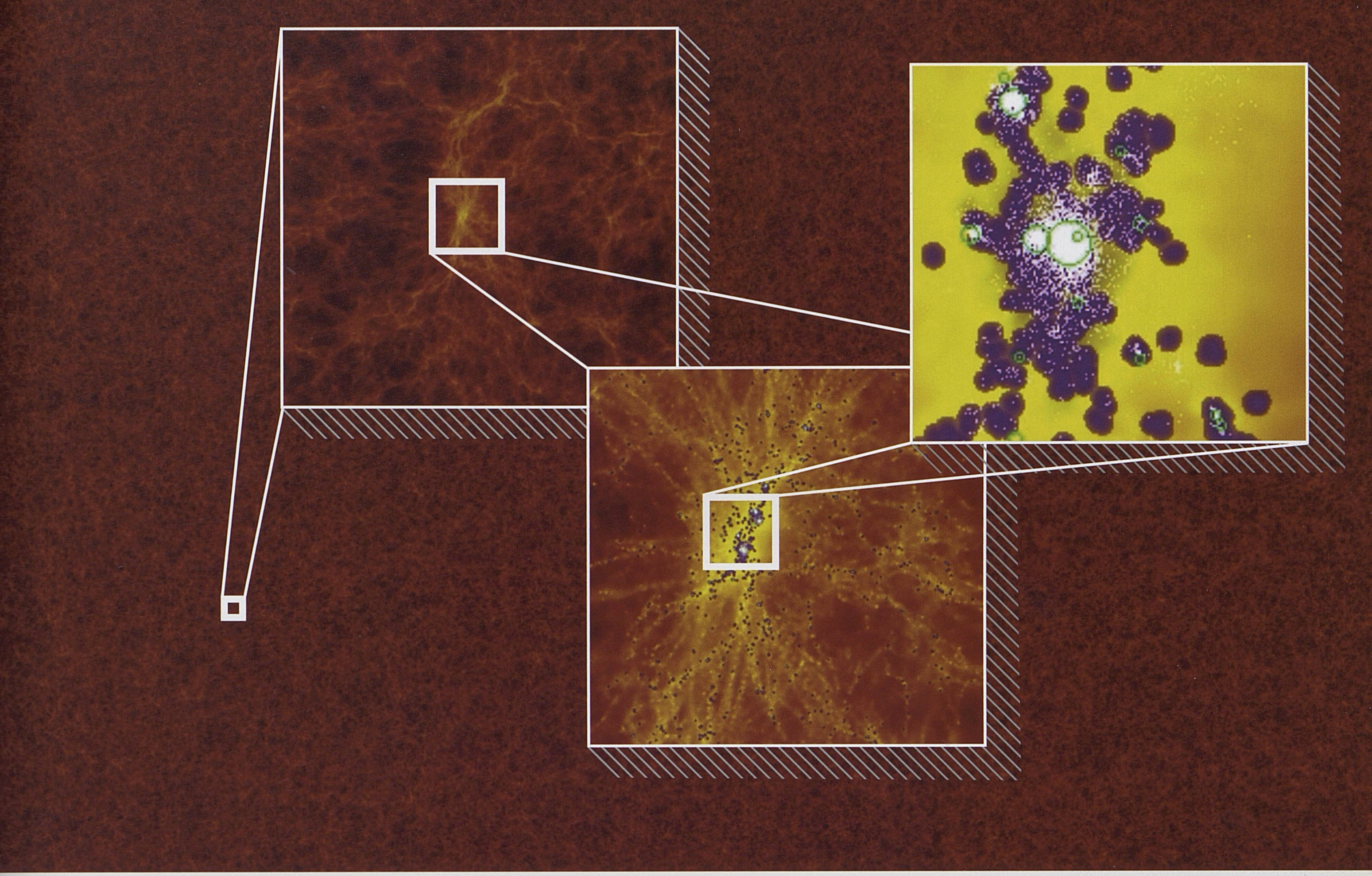
The availability of the powerful Kraken system was crucial, and Di Matteo, Croft and colleagues worked extensively for more than a year to include more physics with GADGET and to optimize it to run efficiently (to "scale") on Kraken's much larger processor count. PSC and XSEDE staff, including XSEDE advanced user support consultant Anirban Jana, helped with testing and benchmarking on various systems. The result with MassiveBlack, say the researchers, was a simulation high enough in resolution to follow how mass is distributed in the inner regions of galaxies, including how stars form and black holes grow as this huge volume of space evolves.

"It provides a unique framework to study the formation of the first quasars," the researchers wrote in their paper (*Astrophysical Journal Letters*, submitted). They saw dense matter forming web-like filaments of structure, a phenomenon seen in the earlier PSC simulation on BigBen. But what the researchers also saw happening in the early universe with MassiveBlack — unavailable to discovery without large-scale simulation — was that these filaments essentially become pipelines of highly dense gas shooting directly to the center of black holes. "Very dense blobs of gas are going straight in," says Di Matteo. "Glug."

SEEING IS BELIEVING

With simulations such as MassiveBlack and others that produce enormous quantities of data, the ability to see simulation data in visual form is a big part of discovery. For MassiveBlack, graduate student Yu Feng, as part of Di Matteo and Croft's team, created innovative visualization tools for which the shared memory of Blacklight, PSC's newest system, was essential.

From the total of 36 snapshots of the simulation volume's evolution, the researchers identified 10 that captured formation of the first quasars and black holes, with each snapshot comprising between three and four terabytes of data. Using PSC's Lustre WAN storage system, the researchers transferred this data from Tennessee to Pittsburgh where they could work with it on Blacklight. Under other circumstances, it would take hours to read three terabytes of data from a hard drive, but with Blacklight's ability to hold this much data, an entire snapshot at once in memory, Feng was able to "raster" each of the snapshots from simulation particles into pixels, so that the results were easily viewable.



GIGAPAN VIEW OF MASSIVE BLACK
These images represent a screen grab from the GigaPan viewer. The background (above) shows the entire simulation volume at redshift 4.75, about 1.3 billion years after the big bang. Respective zooms lead to a region that contains one of the most massive black holes, shown in the final zoom. Color indicates density (increasing from red to yellow to blue to green).

To do this kind of viewing with this amount of data is revolutionary

"With an entire dataset in memory," says Croft, "you can use colors to map properties, such as temperature and density, and you can click and zoom-in, move to a different area. Blacklight is the easiest machine to be able to do this. It allows the most transparent coding to manipulate these large datasets."

For interactive viewing, Feng made one of the snapshots, 65 billion particles, more than a trillion pixels of imaged data, available for interactive viewing (and public access) through the GigaPan web interface: http://www.gigapan.org Developed by Carnegie Mellon in collaboration with NASA Ames Intelligent Robotics Group, with support from Google, GigaPan provides interactive gigapixel viewing on the web. Also for GigaPan viewing, the researchers created an interactive zoomable animation of a smaller simulation — 1,000 frames representing the complete time evolution of early quasar and black hole formation.

These visualization tools, says Di Matteo, were vital to their "cold gas flow" findings, and to do this kind of viewing with this amount of data, she believes, is revolutionary: "You can pan through the entire volume. It's all there. You can look at details, and you can change your mind and look somewhere else and compare. Before you would have to run code to do that. Here we saw this cold stream going in, and our reaction was 'Wow.'"

MORE INFO:
www.psc.edu/science/2011/supermassive

IN PROGRESS

TWISTED ROPES OF SOLAR WIND

With PSC's Blacklight, a team of physicists is visualizing a fundamental phenomenon involved in space weather that can disrupt satellities, spacecraft and power grids on Earth

A shimmering curtain of iridescence in the night sky, the northern lights, *aurora borealis*, northern dawn, and its southern sister, *aurora australis*, rank among mother nature's more awe-inspiring spectacles. Scientists now know that the auroras are a result of solar wind, ionized particles blasted into space by the sun, crashing into Earth's magnetic field.

Some of these same events that create beautiful light shows have a dark side — the ability to play havoc with electronics. They've knocked out satellites and, on occasion, caused power blackouts on Earth. At the detailed level of physics, all these events, including the auroras, occur due to a phenomenon called "magnetic reconnection" — what happens when Earth's magnetic field lines break and reconnect in ways that allow solar particles to penetrate deeply enough to cause trouble.

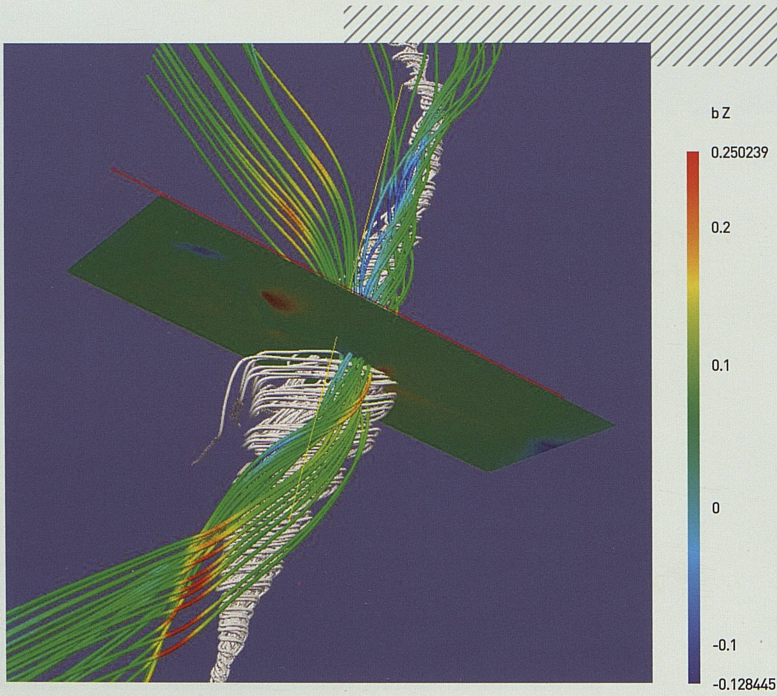
"Magnetic reconnection is a physical process that is prevalent throughout the universe," says physicist Homa Karimabadi at the University of California, San Diego. "It's the predominant mechanism that fractures the Earth's protective magnetic shield exposing us to the effects of solar activity."

Karimabadi works with a team of physicists who have used petascale supercomputing to carry out, for the first time, realistic 3D simulations of magnetic reconnection. With their very large-scale simulations using Kraken, the powerful XSEDE resource at NICS in Tennessee and another system, Karimabadi and his collaborators have been able to characterize, with much greater realism that was previously possible, how turbulence within sheets of electrons generates helical magnetic structures called "flux ropes" — which physicists believe play a large role in magnetic reconnection.

Karimabadi is using PSC's Blacklight to visualize their recent simulations. "Our simulations produce a huge amount of data," says Karimabadi. "One run can generate over 200 terabytes. Blacklight's shared-memory architecture is critical for analysis of these massive data sets." The results of their study are important for NASA's upcoming Magnetosphere Multiscale Mission to observe and measure magnetic reconnection.

FLUX ROPE IN A CURRENT SHEET

This 3D graphic from visualization on Blacklight shows magnetic-field lines (intensity coded by color, blue through red, negative to positive) and associated tornado-like streamlines (white) of a large flux rope formed, notes Karimabadi, due to "tearing instability in thin electron layers." Karimabadi and colleagues reported their findings in *Nature Physics* (April 2011).



SHUTTING THE DOOR ON HIV

With PSC resources, an NIH-funded team is making progress toward finding a therapeutic drug that can deliver a knockout punch to AIDS

AIDS-related research has led to powerful anti-viral drugs that increase life expectancy for people infected with the human immuno-deficiency virus (HIV). Even at their most successful, however, these drugs aren't a cure-all. The United Nations estimates that 33 million people worldwide were living with HIV at the end of 2009, up from 26 million in 1999, and that AIDS in 2009 claimed 1.8 million lives. For researchers, the quest remains not only to find therapeutic agents that manage the disease, but to stop HIV infection before it begins.

To that end, computational chemist Judith LaLonde of Bryn Mawr College works with a multi-faceted team of researchers to develop "inhibitors" — therapeutic drug compounds — that can prevent HIV from gaining entry to cells. "Most HIV therapeutics target replication," says LaLonde, "or integration of the genetic material, but there's very few inhibitors that target the first stage, which is recognition and entry of the virus into the human cell."

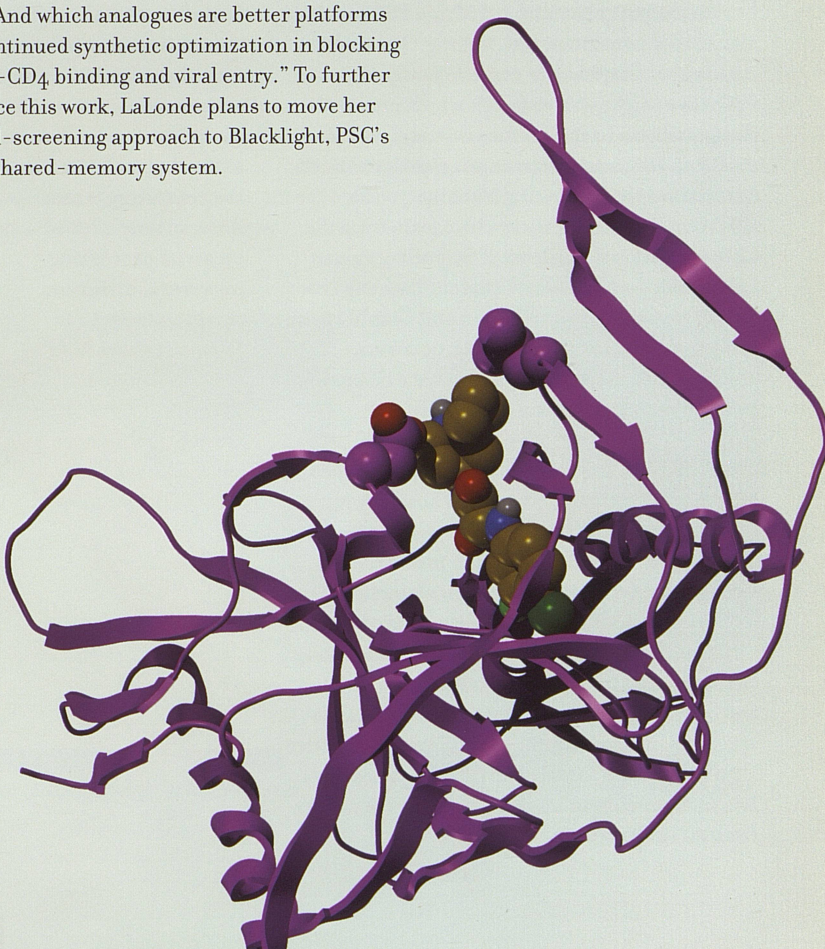
The NIH-funded team includes virologists, chemists and crystallographers, and with LaLonde as a computational specialist, their focus is a protein, called gp120 ("gp" for glycoprotein) on the surface of HIV-1. As part of the initial encounter of HIV with a host cell, gp120 binds to a receptor protein, called CD4, on the host cell's surface. Studies show that a large cavity forms inside gp120 as it binds to CD4. LaLonde's objective is to computationally identify compounds that can bind in that gp120 cavity and prevent it from binding with CD4, shutting the door on infection.

Using PSC's Warhol system, a 64-core Hewlett-Packard cluster, LaLonde runs a "virtual screening" program called ROCS, which uses shape-based similarity matching of small-molecule compounds. Her starting point in recent work was a compound, called NBD-556, shown to have potential as an inhibitor of gp120-CD4 binding. With ROCS and Warhol, she searched for compounds that matched structural features with NBD-556. "With the cluster at PSC," says LaLonde, "I was able to screen eight-million commercially purchasable compounds from the Zinc database in 24 hours."

Her work identified a subset of close matches to NBD-556. With further analysis, the research team narrowed this group of candidates and synthesized those with the most potential. "We discovered new analogues," says LaLonde, "with biological profiles that are improved from where we started. With this larger repertoire, we can ask 'Which compounds work better and why? And which analogues are better platforms for continued synthetic optimization in blocking gp120-CD4 binding and viral entry." To further advance this work, LaLonde plans to move her virtual-screening approach to Blacklight, PSC's large shared-memory system.

STOPPING HIV ENTRY

This image represents the structure of gp120 (purple), a three-part (trimer) HIV protein, with a bridging sheet (left), an outer domain (right), and an inner domain that forms a cavity during binding with host-cell CD4 receptors (not shown), but here is bound with an inhibitor molecule discovered through LaLonde's virtual screening.



JUKEBOX WITH A BRAIN

With PSC's Blacklight on their side, a team of machine learning scientists came in near the front of the pack in a prestigious international competition

Algorithms that allow computers to learn, that is, to change their behavior based on information they encounter, describes the branch of artificial intelligence called "machine learning." The world's first Department of Machine Learning is at Carnegie Mellon University, and post-doctoral researcher Danny Bickson focuses his work there on developing machine-learning software, called GraphLab. A project initiated by Carnegie Mellon professor Carlos Guestrin, GraphLab is open-source software for solving large machine-learning problems with parallel computing. Recently released to the scientific community, it has about 1600 installations around the world.

During the past year, with help from PSC scientist and XSEDE consultant Joel Welling, Bickson customized GraphLab to run efficiently on PSC's Blacklight system (p. 4), which provided computational firepower for an annual, worldwide machine-learning competition, sponsored by the Association for Computing Machinery, called the KDD Cup. In collaboration with a team from the Chinese National Academy of Science — named LeBuSiShu — and graduate student Yucheng Low from Carnegie Mellon, Bickson and GraphLab came in fifth among more than 1,000 teams, ahead of IBM, AT&T and many other well known computer companies and universities.

The competition involved predicting how much people will like songs — based on how they rated other songs. Part of the challenge was the huge dataset — more than 260 million music ratings from the Yahoo music service, which includes 625,000 songs with ratings — on a scale of 1 to 100 — by a million different listeners. Each rated song has associated information — album and artist, and one or more genres. The LeBuSiShu team's approach involved running 12 different predictive algorithms, with a total of 53 tunable parameters. The final prediction involved merging the results from all 12 algorithms.

Another challenge involved the taxonomical relationships among the songs, based on their artist, album and genre connections. Bickson's approach employed a novel method called Matrix Factorization Item Taxonomy Regularization (MFITR). Predictions were tested against a portion of the Yahoo dataset that was held out from the competition, and MFITR by itself produced the second best prediction result among the dozen algorithms Bickson's team ran via the GraphLab framework.

"For each algorithm," says Bickson, "there can be many, many runs — as many as you can — to fine tune different parameters and find what works best. When you have more computing power, you can tune your algorithms faster, and you get better results in the short time frame of the contest." Most of the competing teams, Bickson notes, relied on serial processing on several different machines with large groups of researchers doing the computing. With Blacklight on his team, Bickson and GraphLab more than held their own.

WHAT SHAPES THE WIND?

A team of scientists is using computational modeling to help develop a Climate Action Plan for the Los Angeles region

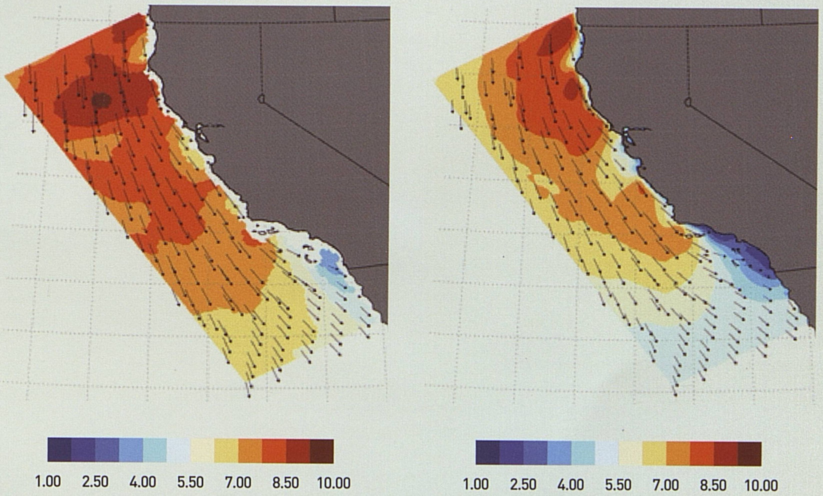
"Los Angeles weather is the weather of catastrophe, of apocalypse, and, just as the reliably long and bitter winters of New England determine the way life is lived there, so the violence and the unpredictability of the Santa Ana affect the entire quality of life in Los Angeles, accentuate its impermanence, its unreliability. The wind shows us how close to the edge we are."

Joan Didion,
Los Angeles Notebook

Alex Hall, a professor in UCLA's Institute of the Environment and Sustainability, is interested in the uncertainties of global climate change and, more specifically, in how climate change may have regional effects. This means thinking about factors not well accounted for in large-scale global climate models. In the Los Angeles area, where Hall lives and works, for instance, it means taking into account how the coastal Pacific ocean affects conditions over the rugged Sierra mountains not far inland — natural features that, among other effects, shape winds, known as the Santa Anas, that have fueled some the most furious wildfires to occur in densely populated areas.

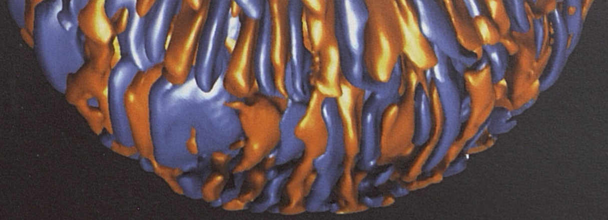
Hall works with a coalition of government, universities and private concerns whose aim is to develop a Climate Action Plan that considers some of the "What if?" questions involved with climate change in the Los Angeles region. "The Santa Ana phenomenon, among other factors, isn't represented at all in the coarse resolution global models," says Hall. "To recover that phenomenon you have to regionalize the simulation at high resolution. Studying the dynamics of these phenomena scientifically, understanding the climate system at these smaller scales, is a very important objective."

Recent results from modeling by Hall and colleagues show that the connection between sea-surface temperature and winds at high resolution must be taken into account in considering the regional effects of climate change. They have used PSC's Blacklight, among other computing resources, and solved a series of challenges in running computational experiments that integrate the effects of the coastal ocean with those of inland topography. At a resolution of two kilometers (compared to the roughly 100-kilometer resolution of global models), their work captures the interplay between surface temperatures of the ocean and inland mountains in creating wind and rain conditions. "In the case of California," says Hall, "we can reproduce rain events when they actually occurred going back as far as data is available." In ongoing work, they expect to investigate some of the uncertainties associated with Los Angeles regional climate change scenarios, including not only the Santa Ana winds, but also critical questions associated with availability of water resources.



SOUTHERN CALIFORNIA WIND IN THE SUMMER OF 2002

The first graphic (left) shows mean wind speed direction (arrows) and magnitude (decreasing from red to blue) measured by satellite compared to the results (right) of their coupled ocean and atmosphere model.



PITTSBURGH
SUPERCOMPUTING
CENTER

The Pittsburgh Supercomputing Center is a joint effort of Carnegie Mellon University and the University of Pittsburgh together with Westinghouse Electric Company. It was established in 1986 and is supported by several federal agencies, the Commonwealth of Pennsylvania and private industry.

PSC GRATEFULLY ACKNOWLEDGES SIGNIFICANT SUPPORT FROM THE FOLLOWING:

The Commonwealth of Pennsylvania
The National Science Foundation
The National Institutes of Health
The National Energy Technology Laboratory
The National Oceanographic and Atmospheric Administration
The National Archives and Records Administration
The U. S. Department of Defense
The U. S. Department of Energy
Cisco Systems, Inc.
Cray Inc.
DataDirect Networks
DSF Charitable Foundation
Microsoft Corporation
Silicon Graphics, Inc.
The Buhl Foundation
Bill and Melinda Gates Foundation

EDITOR/WRITER: Michael Schneider, PSC

CONTRIBUTING WRITER: Deb Nigra, PSC

PROJECT COORDINATION: Shandra Williams, PSC

TRANSCRIBING: Ryan Omecene, PSC & PSC hotline staff

PHOTOGRAPHY: Tim Kaulen & Jordan Bush, Photography & Graphic Services at Mellon Institute.

GRAPHICS: Thanks to the researchers & PSC scientists.

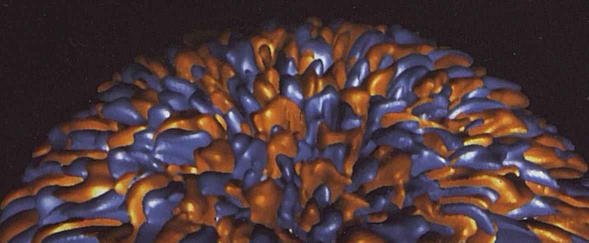
COVER GRAPHIC: The turbulent-convection zone of a young star from a simulation by Ben Brown of the Juri Toomre research group, University of Colorado, on BigBen (PSC) and Kraken (NICS). PSC visualization specialist Greg Foss rendered the image, which shows plasma-flow velocity streamlines (originally blue and gold, to indicate direction), on PSC's Blacklight.

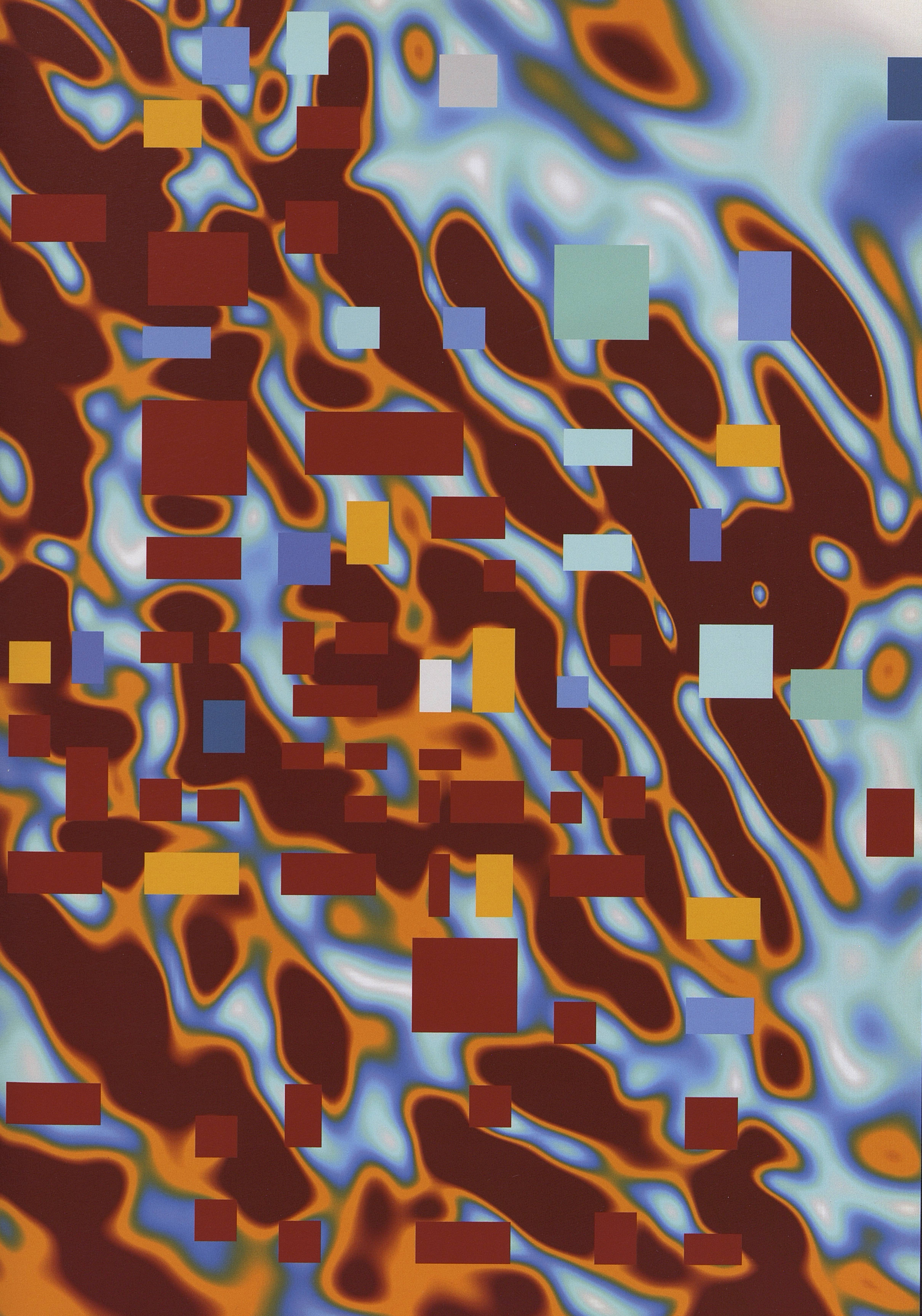
DESIGN: Wall-to-Wall Studios, Inc.

PRINTING: Heeter Direct Printing

FSC®
www.fsc.org
MIX
Paper from responsible sources
FSC® C021348

Printed on Sappi McCoy Paper, a premium sheet with 10 percent post-consumer waste fiber, with vegetable-based inks.





PITTSBURGH SUPERCOMPUTING CENTER
300 S. CRAIG STREET
PITTSBURGH, PENNSYLVANIA 15213