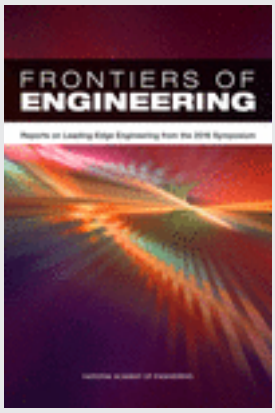

Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2016 Symposium

DETAILS

144 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-45036-2 | DOI: 10.17226/23659

AUTHORS

National Academy of Engineering

FRONTIERS OF ENGINEERING

Reports on Leading-Edge Engineering from the 2016 Symposium

NATIONAL ACADEMY OF ENGINEERING

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS • 500 Fifth Street, NW • Washington, DC 20001

Funding for the activity that led to this publication was provided by The Grainger Foundation, Defense Advanced Research Projects Agency, Department of Defense ASD(R&E) Research Directorate—STEM Development Office, Air Force Office of Scientific Research, Microsoft Research, Cummins Inc., and individual donors. This material is also based on work supported by the National Science Foundation under Grant No. 1611723. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation. In addition, the content of this publication does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

International Standard Book Number-13: 978-0-309-45036-2
International Standard Book Number-10: 0-309-45036-5
Digital Object Identifier: 10.17226/23659

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; Internet, http://www.nap.edu.

Printed in the United States of America

Suggestion citation: National Academy of Engineering. 2017. *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2016 Symposium*. Washington, DC: The National Academies Press. doi: 10.17226/23659.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

ORGANIZING COMMITTEE

ROBERT D. BRAUN (Chair), Dean of Engineering and Applied Science, University of Colorado Boulder
JULIE CHAMPION, Associate Professor, School of Chemical and Biomolecular Engineering, Georgia Institute of Technology
AMY CHILDRESS, Professor, Sonny Astani Department of Civil and Environmental Engineering, University of Southern California
DESHAWN JACKSON, Business Analyst, Production Enhancement, Halliburton
DAVID LUEBKE, Vice President, Graphics Research, NVIDIA
JOHN OWENS, Professor, Department of Electrical and Computer Engineering, University of California, Davis
MARCO PAVONE, Assistant Professor, Department of Aeronautics and Astronautics, Stanford University
ABHISHEK ROY, Senior Research Scientist, Energy and Water Solutions, The Dow Chemical Company
PETER TESSIER, Richard Baruch M.D. Career Development Associate Professor, Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute

Staff

JANET HUNZIKER, Senior Program Officer
SHERRI HUNTER, Program Coordinator

Preface

This volume presents papers on the topics covered at the National Academy of Engineering's 2016 US Frontiers of Engineering Symposium. Every year the symposium brings together 100 outstanding young leaders in engineering to share their cutting-edge research and innovations in selected areas. The 2016 symposium was held September 19–21 at the Arnold and Mabel Beckman Center in Irvine, California. The intent of this book is to convey the excitement of this unique meeting and to highlight innovative developments in engineering research and technical work.

GOALS OF THE FRONTIERS OF ENGINEERING PROGRAM

The practice of engineering is continually changing. Engineers must be able not only to thrive in an environment of rapid technological change and globalization but also to work on interdisciplinary teams. Today's research is being done at the intersections of engineering disciplines, and successful researchers and practitioners must be aware of developments and challenges in areas that may not be familiar to them.

At the annual 2½-day US Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers—ages 30 to 45, from academia, industry, and government and a variety of engineering disciplines—learn from their peers about pioneering work in different areas of engineering. The number of participants is limited to 100 to maximize opportunities for interactions and exchanges among the attendees, who are chosen through a competitive nomination and selection process. The symposium is designed to foster contacts and learning among promising individuals who would not meet in the usual round of professional

meetings. This networking may lead to collaborative work, facilitate the transfer of new techniques and approaches, and produce insights and applications that bolster US innovative capacity.

The four topics and the speakers for each year's meeting are selected by an organizing committee of engineers in the same 30- to 45-year-old cohort as the participants. Speakers describe the challenges they face and communicate the excitement of their work to a technically sophisticated but nonspecialist audience. They provide a brief overview of their field of inquiry; define the frontiers of that field; describe experiments, prototypes, and design studies (completed or in progress) as well as new tools and methods, limitations and controversies; and assess the long-term significance of their work.

The 2016 Symposium

The topics covered at the 2016 symposium were (1) pixels at scale: high-performance computer graphics and vision, (2) extreme engineering: extreme autonomy in space, air, land, and under water, (3) water desalination and purification, and (4) technologies for understanding and treating cancer.

The first session on computer graphics and vision addressed the question, "What do we do with all the pixels brought about by advances in computer graphics hardware, high-resolution displays, and high-resolution, low-cost digital cameras?" The speakers focused on four interrelated technology and application areas: computer vision and image understanding, modern computer graphics hardware, computational display, and virtual reality. The first speaker discussed the relatively new field of computational near-eye display, which operates at the boundary of optics, electronics, and computer graphics to design innovative display systems with new capabilities. This was followed by a talk on pioneering virtual reality headsets, where the display is an inch from the eyes and controlled by one's head and requires performance and resolution significantly beyond what current systems offer. The third speaker covered the pairing of image recognition with learning from that recognition, which has applications in visual search, and first-person vision where the camera wearer is an active participant in visual observation. The session concluded with a presentation on the challenges and opportunities of processing live pixel streams on vast scales with applications ranging from the personal to the societal.

Recent breakthroughs in decision-making, perception architectures, and mechanical design are paving the way for autonomous robotic systems carrying out a wide range of tasks of unprecedented complexity. The session Extreme Engineering: Extreme Autonomy in Space, Air, Land, and Under Water provided an overview of four domains where recent algorithmic and mechanical advances are enabling the design and deployment of robotic systems where autonomy is pushed to the extreme. The session started with a presentation on the challenges of precision landing for reusable rockets, the technology required, and what will be

needed to extend precision landing to planets other than Earth. The next presentation focused on autonomous microflying robots with design innovations inspired by avian flight. This was followed by a talk on the robotic cheetah, the first four-legged robot to run and jump over obstacles autonomously, and the management of balance, energy, and impact without human interaction. The fourth and final presentation covered motion guidance for ocean sampling by underwater vehicles.

Securing a reliable supply of water is a global challenge due to a growing population, changing climate, and increasing urbanization; therefore, alternative sources to augment freshwater supplies are being explored. The third session focused on four critical areas of water desalination and purification: new materials development, analytical characterization techniques, emerging desalination technologies, and innovative system design and operation. The session began with an overview of current reverse osmosis technology, applications, and membrane chemistry innovations, which was followed by a presentation on scalable manufacturing of layer-by-layer membranes and the advanced membrane characterization techniques that drive breakthrough innovations. The third speaker introduced new materials that advance emerging desalination treatment technologies. The final speaker asserted that desalination may present the same challenge for the next 100 years as building the Hoover Dam, which solved water scarcity issues that arose in the 1920s and 1930s. He discussed various high-recovery treatment options that utilize challenging solution chemistries or result in zero liquid discharge.

The organizers of the final session, Technologies for Understanding and Treating Cancer, noted that cancer is a complex group of more than 100 diseases characterized by uncontrolled cell growth, and that approximately 40 percent of people will be diagnosed with a form of cancer in their lifetime. Cancer presents challenges that engineers from different disciplines are working to address, through, for example, the development of more selective tools to detect cancer, new methods to deliver drugs to cancer cells, and better imaging methods to identify smaller tumors and assist surgeons in removing only cancerous cells. The session opened with a talk on how extracellular signals and the microenvironment around cancer cells influence their uncontrolled growth and expansion. This was followed by a presentation on advances in noninvasive methods using microfluidics to detect rare cancer cells. The third speaker described therapeutic molecules that block the ability of cancer cells to leave the initial tumor and start new tumors. The last speaker talked about immunotherapy—strategies for harnessing the immune system to target cancer cells using methods that control and sustain anti-tumor immune responses specific for different types of cancer.

In addition to the plenary sessions, the attendees had many opportunities for informal interaction. On the first afternoon, they gathered in small groups for "get-acquainted" sessions during which they presented short descriptions of their work and answered questions from their colleagues. This helped them get to know more about each other relatively early in the program. On the second

afternoon, attendees met in small groups to discuss issues such as inspiring and training (from K through PhD) future engineering leaders, industry-academic-government collaboration, sustainable energy systems, wearable technology, and change management in industries and disciplines where technology is rapidly improving, among others.

Every year a distinguished engineer addresses the participants at dinner on the first evening of the symposium. The 2016 speaker, NAE member John A. Orcutt, distinguished professor of geophysics at the Scripps Institution of Oceanography and the University of California, San Diego, gave the first evening's dinner speech titled, "The Arctic: Scientific and Engineering Challenges for Measuring Rapid Change." He made a compelling case for climate research by enumerating significant Pan-Arctic changes—reduction in sea-ice thickness, warming of Arctic waters and permafrost, rising temperatures, melting of the Greenland Ice Sheet, and increase in human activities as well as economic and geopolitical importance—resulting from climate change. He described the sensing networks such as Arctic Watch that employ communication, underwater navigation, and acoustic remote sensing technologies to observe, monitor, and collect data in situ year-around.

The NAE is deeply grateful to the following for their support of the 2016 US Frontiers of Engineering symposium:

- The Grainger Foundation
- Defense Advanced Research Projects Agency
- Air Force Office of Scientific Research
- Department of Defense ASD(R&E)–STEM Development Office
- National Science Foundation (this material is based on work supported by the NSF under grant EFMA-1611723)
- Microsoft Research
- Cummins Inc.
- Individual contributors

We also thank the members of the Symposium Organizing Committee (p. iv), chaired by Dr. Robert Braun, for planning and organizing the event.

Contents

APPENDIXES

Pixels at Scale: High-Performance Computer Graphics and Vision

Pixels at Scale: High-Performance Computer Graphics and Vision

David Luebke
NVIDIA Research

John Owens
University of California, Davis

The smartphones we carry in our pockets have remarkable capabilities that were unimaginable only a decade ago: a high-quality retinal display powered by high-performance graphics hardware, a high-resolution camera capable of capturing a billion pixels every second, and a high-bandwidth connection to a cloud infrastructure with tremendous computational horsepower. In short, we now have an abundance of pixels that can be produced, processed, and consumed easily and cheaply. This session addresses the question, What do we do with all these pixels?

The ascendance of the pixel is the culmination of numerous technical advances that began with the invention of computer graphics in the 1960s and digital photography in the 1970s. Modern consumer hardware has made ubiquitous

- powerful computer graphics hardware, continuously increasing the performance and quality of graphics;
- high-resolution displays, approaching the native resolution of the eye; and
- high-resolution low-cost digital cameras, generating trillions of digital photos for analysis and training.

These advances provide traction on two long-standing challenges that center on the pixel: interactive, immersive, photorealistic computer graphics and ubiquitous, robust image analysis and understanding. Speakers in this session discussed four interlocking technology and application areas spanning "pixels in" and "pixels out": computer vision and image understanding, modern computer graphics hardware, computational display, and virtual reality.

Fueled in part by the deluge of pixels—the availability of images for train-

ing at massive scale—advances in machine learning are bringing about a sort of Golden Age of computer vision. Many challenges in machine learning, such as outperforming humans at recognizing objects or understanding speech, have fallen. Computer vision researchers can now tackle problems and applications of image understanding that were previously hard to imagine.

Computer graphics hardware is the computational substrate for the pixel revolution. The graphics processing unit (GPU) in today's PCs and smartphones represents decades of coevolution between graphics algorithms and the silicon architectures that execute them. In the process the modern GPU has grown from a fixed-function coprocessor to a general-purpose parallel computing platform—and accrued considerably more computational horsepower than the rest of the processors in the device put together.

The GPUs on which consumers play video games execute tens of thousands of concurrent threads, providing a level of massively parallel computation that was once the exclusive preserve of supercomputers. Thus today's GPUs not only render video games but also accelerate computation for astrophysics, video transcoding, image processing, protein folding, seismic exploration, computational finance, heart surgery, self-driving cars—the list goes on and on. Importantly, machine learning algorithms (particularly convolutional neural nets or "deep learning") map especially well to GPUs, which largely power the computer vision renaissance.

Pixels are just data until a display turns them into photons, and display technology is undergoing its own tectonic shifts. LCD and OLED panels are following their own Moore's Law and achieving breathtaking advances in resolution, cost, size (both large and small), and brightness—almost every metric one can think of.

Less obvious is a body of work in computational display, which codesigns the optics and electronics of the display system with the rendering algorithms that generate the pixels. For example, stacking multiple panels can create a light field display, providing glasses-free "3D" (stereo) views with correct motion parallax as the viewer moves. Other novel optics coupled with rendering algorithms enable new tradeoffs, such as trading resolution for a thinner display or focus cues.

Right now, perhaps the most talked-about applications in the ongoing pixel revolution are virtual and augmented reality. We are captivated by the concept of rendering a virtual world so effectively that the perceptual system accepts it as reality, or by the prospect of seamlessly integrating synthetic information and objects into our view of the real world. Virtual and augmented reality pose huge challenges for all the topics discussed above: computer vision must track the user's slightest motions and gestures and interpret their environment; graphics hardware must render at unprecedented levels of performance to achieve immersion; and displays must evolve from today's boxy headmounts to something as vanishingly unobtrusive as a pair of eyeglasses.

The session began with Gordon Wetzstein of Stanford University. Prof. Wetzstein has pioneered the relatively young field of computational display,

working on the boundary of optics, electronics, and computer graphics to design innovative display systems with entirely new capabilities.

Next we welcomed Warren Hunt of Oculus. Oculus, now owned by Facebook, is pioneering virtual reality headsets. Dr. Hunt's (and Oculus's) work is fascinating for two reasons: (1) traditional assumptions in computer graphics are upended when the display is an inch from your eye and controlled by your head; and (2) immersive virtual reality requires performance and resolution significantly beyond what current systems offer.

We then heard from Kristen Grauman, associate professor of computer science at the University of Texas and an expert in computer vision, with particular expertise in the interface between vision and machine learning. Her research couples image recognition with learning from that recognition, with applications in visual search, and a recent focus on first-person vision (enabled, in turn, by advances in cameras) where the camera wearer is an active participant in visual observation.

The session concluded with Kayvon Fatahalian, an assistant professor of computer science at Carnegie Mellon University whose research couples a systems mindset with deep expertise in pixel-processing hardware and software. He discussed the challenges and opportunities of processing live pixel streams on vast scales, with applications ranging from personal to urban to societal.

Computational Near-Eye Displays: Engineering the Interface to the Digital World

Gordon Wetzstein
Stanford University

Immersive virtual reality and augmented reality (VR/AR) systems are entering the consumer market and have the potential to profoundly impact society. Applications of these systems range from communication, entertainment, education, collaborative work, simulation, and training to telesurgery, phobia treatment, and basic vision research. In every immersive experience, the primary interface between the user and the digital world is the near-eye display. Thus, developing near-eye display systems that provide a high-quality user experience is of the utmost importance.

Many characteristics of near-eye displays that define the quality of an experience, such as resolution, refresh rate, contrast, and field of view, have been significantly improved in recent years. However, a significant source of visual discomfort prevails: the vergence-accommodation conflict (VAC), which results from the fact that vergence cues (e.g., the relative rotation of the eyeballs in their sockets), but not focus cues (e.g., deformation of the crystalline lenses in the eyes), are simulated in near-eye display systems. Indeed, natural focus cues are not supported by any existing near-eye display.

Using focus-tunable optics, we explore unprecedented display modes that tackle this issue in multiple ways with the goal of increasing visual comfort and providing more realistic visual experiences.

INTRODUCTION

In current VR/AR systems, a stereoscopic near-eye display presents two different images to the viewer's left and right eyes. Because each eye sees a slightly different view of the virtual world, binocular disparity cues are created

that generate a vivid sense of three-dimensionality. These disparity cues also drive viewers' vergence state as they look around at objects with different depths in the virtual world.

However, in a VR system the accommodation, or focus state, of the viewer's eyes is optically fixed to one specific distance. This is because, despite the simulated disparity cues, the micro display in a VR system is actually at a single, fixed optical distance. The specific distance is defined by the magnified image of the micro display, and the eyes are forced to focus at that distance and only that distance in order for the virtual world to appear sharp. Focusing at other distances (such as those simulated by stereoscopic views) results in a blurred view.

In the physical world, these two properties of the visual response—vergence and accommodation—work in harmony (see Figure 1). Thus, the neural systems that drive the vergence and accommodative states of the eye are neurally coupled.

VR/AR displays artificially decouple vergence and focus cues because their image formation keeps the focus at a fixed optical distance but drives vergences to arbitrary distances via computer-generated stereoscopic imagery. The resulting discrepancy—the vergence-accommodation conflict—between natural depth cues and those produced by VR/AR displays may lead to visual discomfort and fatigue, eyestrain, double vision, headaches, nausea, compromised image quality, and even pathologies in the developing visual system of children.

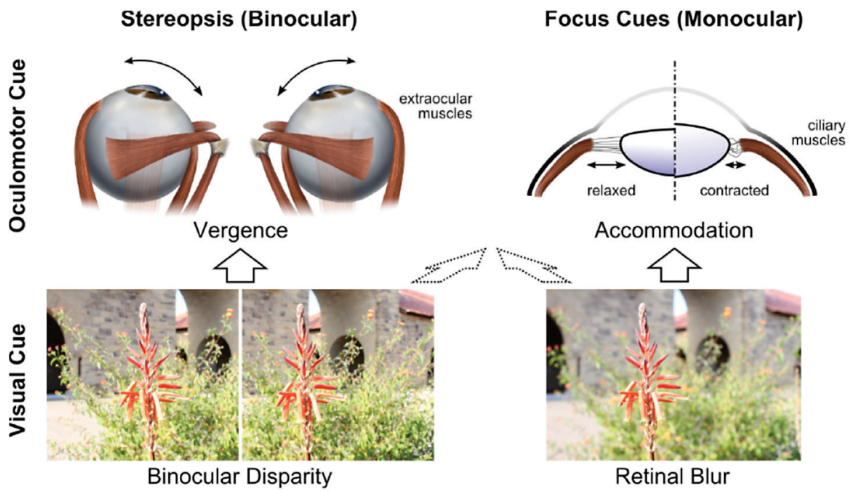

FIGURE 1 Overview of relevant depth cues. Vergence and accommodation are oculomotor cues; binocular disparity and retinal blur are visual cues. In normal viewing conditions, disparity drives vergence and blur drives accommodation. However, these cues are cross-coupled. Near-eye displays support only binocular cues, not focus cues.

The benefits of providing correct or nearly correct focus cues include not only increased visual comfort but also improvements in 3D shape perception, stereoscopic correspondence matching, and discrimination of larger depth intervals. Significant efforts have therefore been made to engineer focus-supporting displays.

But all technologies that might support focus cues suffer from undesirable tradeoffs in compromised image resolution, device form factor or size, and brightness, contrast, or other important display characteristics. These tradeoffs pose substantial challenges for high-quality AR/VR visual imagery with practical, wearable displays.

BACKGROUND

In recent years a number of near-eye displays have been proposed that support focus cues. Generally, these displays can be divided into the following classes: adaptive focus, volumetric, light field, and holographic displays.

Two-dimensional adaptive focus displays do not produce correct focus cues: the virtual image of a single display plane is presented to each eye, just as in conventional near-eye displays. However, the system is capable of dynamically adjusting the distance of the observed image, either by actuating (physically moving) the screen (Sugihara and Miyasoto 1998) or using focus-tunable optics (programmable liquid lenses). Because this technology only enables the distance of the entire virtual image to be adjusted at once, the correct focal distance at which to place the display will depend on where in the simulated 3D scene the user is looking.

Peli (1999) reviews several studies that proposed the idea of gaze-contingent focus, but I am not aware of anyone having built a practical gaze-contingent, focus-tunable display prototype. The challenge for this technology is to engineer a robust gaze and vergence tracking system in a head-mounted display with custom optics.

A software-only alternative to gaze-contingent focus is gaze-contingent blur rendering (Mauderer et al. 2014), but because the distance to the display is still fixed in this technique it does not affect the VAC. Konrad and colleagues (2016) recently evaluated several focus-tunable display modes in near-eye displays and proposed monovision as a practical alternative to gaze-contingent focus, where each eye is optically accommodated at a different depth.

Three-dimensional volumetric and *multiplane displays* represent the most common approach to focus-supporting near-eye displays. Instead of using 2D display primitives at a fixed or adaptive distance to the eye, volumetric displays either mechanically or optically scan the 3D space of possible light-emitting display primitives (i.e., pixels) in front of each eye (Schowengerdt and Seibel 2006).

Multiplane displays approximate this volume using a few virtual planes generated by beam splitters (Akeley et al. 2004; Dolgoff 1997) or time-multiplexed focus-tunable optics (Liu et al. 2008; Llull et al. 2015; Love et al. 2009; Rolland et al. 2000; von Waldkirch et al. 2004). Whereas implementations with beam

splitters compromise the form factor of a near-eye display, temporal multiplexing introduces perceived flicker and requires display refresh rates beyond those offered by current-generation microdisplays.

Four-dimensional light field and *holographic displays* aim to synthesize the full 4D light field in front of each eye. Conceptually, this approach allows for parallax over the entire eyebox to be accurately reproduced, including monocular occlusions, specular highlights, and other effects that cannot be reproduced by volumetric displays. Current-generation light field displays provide limited resolution (Hua and Javidi 2014; Huang et al. 2015; Lanman and Luebke 2013), whereas holographic displays suffer from speckle and require display pixel sizes to be in the order of the wavelength of light, which currently cannot be achieved at high resolution for near-eye displays, where the screen is magnified to provide a large field of view.

EMERGING COMPUTATIONAL NEAR-EYE DISPLAY SYSTEMS

In our work, we ask whether it is possible to provide natural focus cues and to mitigate visual discomfort using focus-tunable optics, i.e., programmable liquid lenses. For this purpose, we demonstrate a prototype focus-tunable near-eye display system (Figure 2) that allows us to evaluate several advanced display modes via user studies.

Conventional near-eye displays are simple magnifiers that enlarge the image of a microdisplay and create a virtual image at some fixed distance to the viewer.

Adaptive depth of field rendering is a software-only approach that renders the fixated object sharply while blurring other objects according to their relative distance. When combined with eye tracking, this mode is known as gaze-contingent retinal blur (Mauderer et al. 2014). Because the human accommodation system may be driven by the accommodation-dependent blur gradient, this display mode does not reproduce a physically correct stimulus.

Adaptive focus display is a software/hardware approach that either changes the focal length of the lenses or the distance between the micro display and the lenses (Konrad et al. 2016). When combined with eye tracking, this mode is known as gaze-contingent focus. In this mode, the magnified virtual image observed by the viewer can be dynamically placed at arbitrary distances, for example at the distance where the viewer is verged (requires vergence tracking) or at the depth corresponding to their gaze direction (requires gaze tracking). No eye tracking is necessary to evaluate this mode when the viewer is asked to fixate on a specific object, for example one that moves.

Monovision is a common treatment for presbyopia, a condition that often occurs with age in which people lose the ability to focus their eyes on nearby objects. It entails placing lenses with different prescription values for each eye such that one eye dominates for distance vision and the other for near vision.

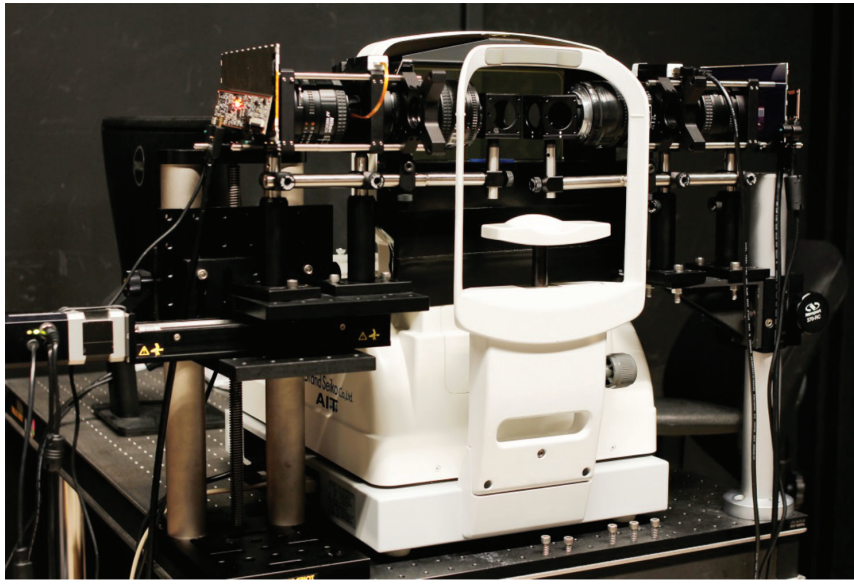

FIGURE 2 Prototype focus-tunable stereoscopic display. This setup allows for a range of focus-tunable and monovision display modes to be tested in user studies. An autorefractor is integrated in the setup to measure where a user accommodates for a displayed stimulus. The outcome of these studies will inform the design of future near-eye displays.

Monovision was recently proposed and evaluated for emmetropic viewers (those with normal or corrected vision) in VR/AR applications (Konrad et al. 2016).

HOW OUR RESEARCH INFORMS NEXT-GENERATION VR/AR DISPLAYS

Preliminary data recorded for our study suggest that both the focus-tunable mode and the monovision mode could improve conventional displays, but both require optical changes to existing VR/AR displays. A software-only solution (i.e., depth of field rendering) proved ineffective. The focus-tunable mode provided the best gain over conventional VR/AR displays. We implemented this display mode with focus-tunable optics, but it could also be implemented by actuating the microdisplay in the VR/AR headset.

Based on our study, we conclude that the adaptive focus display mode seems to be the most promising direction for future display designs. Dynamically changing the accommodation plane depending on the user's gaze direction could improve visual comfort and realism in immersive VR/AR applications in a significant way.

Eye conditions, including myopia (near-sightedness) and hyperopia (far-sightedness), have to be corrected adequately with the near-eye display, so the user's prescription must be known or measured. Presbyopic users cannot accommodate, so dynamically changing the accommodation plane would almost certainly always create a worse experience than the conventional display mode. For them it is crucial for the display to present a sharp image within the user's accommodation range.

In summary, a personalized VR/AR experience that adapts to the user, whether emmetropic, myopic, hyperopic, or presbyopic, is crucial to deliver the best possible experience.

REFERENCES

Akeley K, Watt S, Girshick A, Banks M. 2004. A stereo display prototype with multiple focal distances. ACM Transactions on Graphics (Proc. SIGGRAPH) 23(3):804–813.

Dolgoff E. 1997. Real-depth imaging: A new 3D imaging technology with inexpensive direct-view (no glasses) video and other applications. SPIE Proceedings 3012:282–288.

Hua H, Javidi B. 2014. A 3D integral imaging optical see-through head mounted display. Optics Express 22(11):13484–13491.

Huang FC, Chen K, Wetzstein G. 2015. The light field stereoscope: Immersive computer graphics via factored near-eye light field display with focus cues. ACM Transactions on Graphics (Proc. SIGGRAPH) 34(4):60:1–12.

Konrad R, Cooper E, Wetzstein G. 2016. Novel optical configurations for virtual reality: Evaluating user preference and performance with focus. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1211–1220.

Lanman D, Luebke D. 2013. Near-eye light field displays. ACM Transactions on Graphics (Proc. SIGGRAPH Asia) 32(6):220:1–10.

Liu S, Cheng D, Hua H. 2008. An optical see-through head mounted display with addressable focal planes. Proceedings of the 7th International Symposium on Mixed and Augmented Reality (ISMAR), September 15–18, Cambridge, UK.

Llull P, Bedard N, Wu W, Tosic I, Berkner K, Balram N. 2015. Design and optimization of a near-eye multifocus display system for augmented reality. OSA Imaging and Applied Optics, paper JTH3A.5.

Love G, Hoffman D, Hands D, Gao J, Kirby J, Banks M. 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. Optics Express 17(18):15716–15725.

Mauderer M, Conte S, Nacenta M, Vishwanath D. 2014. Depth perception with gaze-contingent depth of field. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 217–226.

Peli E. 1999. Optometric and perceptual issues with head-mounted displays. In: Visual Instrumentation: Optical Design and Engineering Principles, ed. Mouroulis P. New York: McGraw-Hill.

Rolland J, Krueger J, Goon A. 2000. Multifocal planes head-mounted display. Applied Optics 39(19):3209–3215.

Schowengerdt B, Seibel E. 2006. True 3-D scanned voxel display using single or multiple light sources. Journal of the Society of Information Displays 14(2):135–143.

Sugihara T, Miyasato T. 1998. A lightweight 3-D HMD with accommodative compensation. SID Digest 29(1):927–930.

von Waldkirch M, Lukowicz P, Tröster G. 2004. Multiple imaging technique for extending depth of focus in retinal displays. Optics Express 12(25):6350–6365.

Frontiers in Virtual Reality Headsets

Warren Hunt
Oculus Research

Technological advances are transforming virtual and augmented reality from science fiction to consumer products. When widely deployed, these technologies have the potential for major impact on entertainment, culture, and commerce. This article provides a basic overview of virtual (VR) and augmented (AR) realities, describes some potential high-impact applications, discusses the effort required to achieve these technologies, and explains an aspect of the human visual system that presents challenges for augmented and virtual reality.

WHAT IS VIRTUAL REALITY?

Virtual reality is any simulation created by a computer, presented to a person, and perceivable as real. Current VR devices consist of a head-mounted display, often with headphones. These devices block out a person's hearing and vision of the real world; ultimately, VR technology will encompass more senses.

Although it's not yet possible to display real-time VR content that's indistinguishable from reality, it is already possible to produce an experience referred to as "presence"—the sense that a person has, in fact, been transported somewhere else. This transportation isn't always conscious: a person experiencing presence may logically know they're wearing a headset, but have measurable reactions to virtual objects or threats, such as fear in response to virtual heights.

A variety of companies sell commercial head-mounted displays. Several, such as Sony, have announced future products, and a number of startups are producing prototypes.

GOING BEYOND: AUGMENTED REALITY

While virtual reality aims to completely override human senses, augmented reality systems aim to combine both real-world and virtual stimuli. Devices such as Microsoft's HoloLens use a see-through display to overlay a virtual world onto the real world.

The blending of real and virtual content can augment everything from daily life, such as virtual name tags or line-item reviews superimposed on a restaurant menu, to complex specialized tasks, such as the overlay of MRI data directly onto a patient in an operating room.

WHY VIRTUAL AND AUGMENTED REALITY?

Virtual and augmented reality have a large range of potential uses. The most obvious and immediate are currently commercially available: entertainment, movies, 360 video, and games. These technologies also have numerous additional applications, including social and business communication, journalism, e-commerce, and education.

Virtual reality has been a futuristic technology for a long time now, and many factors suggest that it is now ready to succeed in the main stream. Moore's law has enabled powerful graphics hardware to render high-definition resolutions at frame rates sufficient for a commercially viable visual experience. Moreover, the rise of smartphones has made very high density organic LED (OLED) display panels and low-latency accelerometers, two key components in high-quality VR, inexpensive and widely available.

WHAT DOES IT TAKE TO MAKE VR/AR?

Building a virtual or augmented reality system is a massive multidisciplinary effort. At the heart of this effort are perceptual scientists: they define the requirements for matching and driving the human perceptual system in order to make VR believable and prevent users from experiencing motion sickness or any other form of discomfort from the VR experience. Such requirements include audio/visual fidelity, latency limits, and tracking accuracy, among others.

Building a head-mounted display requires optical, electrical, and mechanical engineering, understanding of displays and tracking technology, and software expertise in graphics, sound, computer vision, and user interaction. These components must be implemented with a great degree of care and coordination.

Furthermore, achieving high quality in graphical and computer vision systems requires extreme amounts of computing power. Virtual and augmented reality systems currently produce a rather crude representation of the world and a resolution far inferior to what humans can perceive. Achieving a virtual system

that is indistinguishable from reality could consume many orders of magnitude more computing power.

A STABLE VIRTUAL WORLD

To achieve a compelling VR experience—and one that minimizes the risk of motion sickness—the virtual world must consistently appear stable to the user. While traditional displays, e.g., a TV or desktop monitor, tend to stay in one place, VR head-mounted displays are worn on a user's head and often move very quickly. This rapid display movement can cause artifacts that can break the sense of immersion or, worse, make the user physically uncomfortable. As with most of the requirements for VR headsets, stability requirements are driven by the human perceptual system.

Vestibulo-Ocular Reflex

The human visual system has one of the fastest reflexes in the human body. The vestibulo-ocular reflex (VOR) stabilizes human vision during head motion, and does so with a latency of 3 neurons (about 10 ms). This reflex is responsible for turning the eyes to compensate for head motion and provide a stable retinal image during typical head movement.

Because of this reflex, the human visual system expects a stable, crisp image during head rotation even when the screen (attached to the head) is moving at 300 degrees per second. The reflex actively destabilizes the view presented on a head-mounted display, causing it to slide across the retina and blur as VOR makes the eye counterrotate.

Visual Artifacts from Displays

Judder

Judder is the blurring effect caused by the VOR and a static head-mounted display. Most displays illuminate pixels for the duration of a frame (about 17 ms at 60Hz or 11 ms at 90Hz). When a user's head turns at 300 degrees/second, 11 ms corresponds to 3.3 degrees, and during this movement the pixel becomes smeared across that angle. To address this, VR displays use "low persistence" mode and are activated for only 1–2 ms out of each frame, displaying black the remainder of the time. This prevents the smearing artifact, but leads to a dimmer display and, under certain conditions, a strobe effect.

Latency

Display latency can cause severe artifacts. Even at a modest 100 degrees per second, a system latency of 30 ms would cause the world to lag by a full 3 degrees behind a viewer's gaze. The constant lag causes a noticeable "swimming" artifact that can be disturbing and lead to motion sickness. Better algorithms, advances in graphics hardware, and careful orchestration between applications and display hardware can reduce these effects.

Rolling

A rolling display illuminates pixels as they arrive over the wire, rather than all at the same time. Old CRT TVs and most OLED phones have rolling displays: they start by illuminating the top row of the screen, then the next, and so on, until the whole screen has been illuminated. Alternatively, a global display reads the entire frame before displaying every pixel simultaneously.

Each approach has pros and cons. The global display adds significant latency: rather than displaying pixels immediately, all pixels wait until the last one to arrive before illuminating. With rolling displays, if users move their eyes during the display update, the image appears to distort or shear depending on the direction of movement. Compensation for this artifact requires the integration of high-quality eye tracking.

Achieving the latency reduction allowed by rolling displays while minimizing artifacts is an open research problem.

CONCLUSION

Virtual and augmented reality are nascent technologies, but have the promise of dramatic worldwide impact. Continued improvements to displays, graphics, and tracking, coupled with enhanced understanding of the human perceptual system, will enable the realization of more applications. A variety of companies are investing heavily in this space in anticipation of its potential impact.

A proper mix of technology and funding is shaping up to make for a very exciting future for virtual reality!

First-Person Computational Vision

KRISTEN GRAUMAN
University of Texas at Austin

Recent advances in sensor miniaturization, low-power computing, and battery life have carved the path for the first generation of mainstream wearable cameras. Images and video captured by a first-person (wearable) camera differ in important ways from third-person visual data. A traditional third-person camera passively watches the world, typically from a stationary position. In contrast, a first-person camera is inherently linked to the ongoing experiences of its wearer—it encounters the visual world in the context of the wearer's physical activity, behavior, and goals.

To grasp this difference concretely, imagine two ways you could observe a scene in a shopping mall. In the first, you watch a surveillance camera video and see shoppers occasionally pass by the field of view of the camera. In the second, you watch the video captured by a shopper's head-mounted camera as he actively navigates the mall—going in and out of stores, touching certain objects, moving his head to read signs or look for a friend. While both cases represent similar situations—and indeed the same physical environment—the latter highlights the striking difference in capturing the visual experience from the point of view of the camera wearer.

This distinction has intriguing implications for computer vision research—the realm of artificial intelligence and machine learning that aims to automate *visual intelligence* so that computers can "understand" the semantics and geometry embedded in images and video.

EMERGING APPLICATIONS FOR FIRST-PERSON COMPUTATIONAL VISION

First-person computational vision is poised to enable a class of new applications in domains ranging well beyond augmented reality to behavior assessment, perceptual mobile robotics, video indexing for life-loggers or law enforcement, and even the quantitative study of infant motor and linguistic development.

What's more, the first-person perspective in computational vision has the potential to transform the basic research agenda of computer vision as a field: from one focused on "disembodied" static images, heavily supervised machine learning for closed-world tasks, and stationary testbeds—to one that instead encompasses embodied learning procedures, unsupervised learning and open-world tasks, and dynamic testbeds that change as a function of the system's own actions and decisions.

My group's recent work explores first-person computational vision on two main fronts:

- **Embodied visual representation learning.** How do visual observations from a first-person camera relate to its 3D ego-motion? What can a vision system learn simply by moving around and looking, if it is cognizant of its own ego-motion? How should an agent—whether a human wearer, an autonomous vehicle, or a robot—choose to move, so as to most efficiently resolve ambiguity about a recognition task? These questions have interesting implications for modern visual recognition problems and representation learning challenges underlying many tasks in computer vision.
- **Egocentric summarization.** An always-on first-person camera is a double-edged sword: the entire visual experience is retained without any active control by the wearer, but the entire visual experience is not substantive. How can a system automatically summarize a long egocentric video, pulling out the most important parts to construct a visual index of all significant events? What attention cues does a first-person video reveal, and when was the camera wearer engaged with the environment? Could an intelligent first-person camera predict when it is even a good moment to take photos or video? These questions lead to applications in personal video summarization, sharing first-person experiences, and in situ attention analysis.

Throughout these two research threads, our work is driven by the notion that the camera wearer is an active participant in the visual observations received. We consider egocentric or first-person cameras of varying sources—those worn by people as well as autonomous vehicles and mobile robots.

EMBODIED VISUAL LEARNING: HOW DOES EGO-MOTION SHAPE VISUAL LEARNING AND ACTION?

Cognitive science indicates that proper development of visual perception requires internalizing the link between "how I move" and "what I see." For example, in their famous "kitten carousel" experiment, Held and Hein (1963) examined how the visual development of kittens is shaped by their self-awareness and control (or lack thereof) of their own physical motion.

However, today's best computer vision algorithms, particularly those tackling recognition tasks, are deprived of this link, learning solely from batches of images downloaded from the Web and labeled by human annotators. We argue that such "disembodied" image collections, though clearly valuable when collected at scale, deprive feature learning methods from the informative physical context of the original visual experience (Figure 1).

We propose to develop *embodied visual representations* that explicitly link what is seen to how the sensor is moving. To this end, we present a deep feature learning approach that embeds information not only from the video stream the observer sees but also from the motor actions he simultaneously makes (Jayaraman and Grauman 2015). Specifically, we require that the features learned in a convolutional neural network exhibit *equivariance,* i.e., respond predictably to transformations associated with distinct ego-motions.

During training, the input image sequences are accompanied by a synchronized stream of ego-motor sensor readings. However, they need not possess any semantic labels. The ego-motor signal could correspond, for example, to the

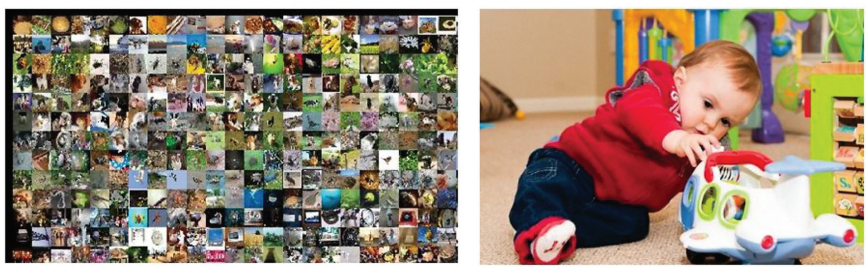

FIGURE 1 (a) The status quo in computer vision is to learn object categories from massive collections of "disembodied" Web photos that have been labeled by human supervisors as to their contents. (b) In first-person vision, it is possible to learn from embodied spatiotemporal observations, capturing not only what is seen but also how it relates to the movement and actions of the self (i.e., the egocentric camera) in the world. Left image is from the ImageNet dataset (Deng et al. 2009); right image is shared by user Daniel under the Creative Commons license.

inertial sensor measurements received alongside video on a wearable or car-mounted camera.

The objective is to learn a function mapping from pixels in a video frame to a space that is equivariant to various motion classes. In other words, the resulting learned features should change in predictable and systematic ways as a function of the transformation applied to the original input (Figure 2).

To exploit the features for recognition, we augment the neural network with a classification loss when class-labeled images are available, driving the system to discover a representation that is also suited for the recognition task at hand. In this way, ego-motion serves as side information to regularize the features learned, which we show facilitates category learning when labeled examples are scarce. We demonstrate the impact for recognition, including a scenario where features learned from "ego-video" on an autonomous car substantially improve large-scale scene recognition.

Building on this concept, we further explore how the system can actively choose *how to move* about a scene, or *how to manipulate* an object, so as to recognize its surroundings using the fewest possible observations (Jayaraman and Grauman 2016). The goal is to learn how the system should move to improve its sequence of observations, and how a sequence of future observations is likely to change conditioned on its possible actions.

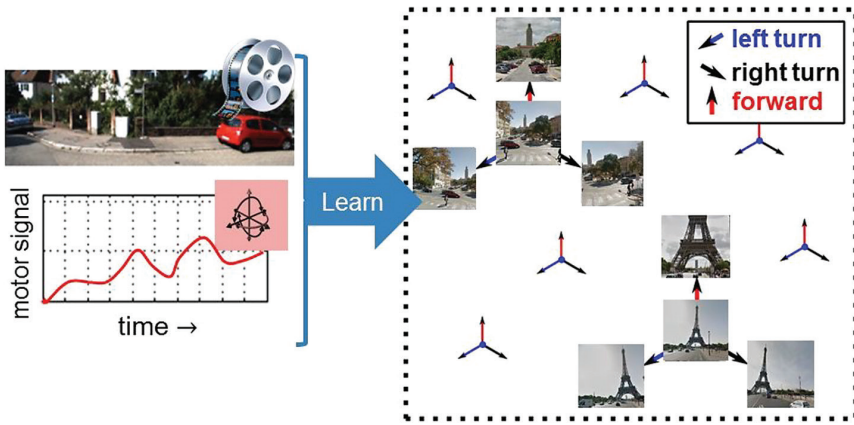

FIGURE 2 Overview of idea to learn visual representations that are equivariant with respect to the camera's ego-motion. Given an unlabeled video accompanied by external measurements of the camera's motion (left), the approach optimizes an embedding that keeps pairs of views organized according to the ego-motion that separates them (right). In other words, the embedding requires that pairs of frames that share an ego-motion be related by the same transformation in the learned feature space. Such a learned representation injects the embodied knowledge of self-motion into the description of what is seen.

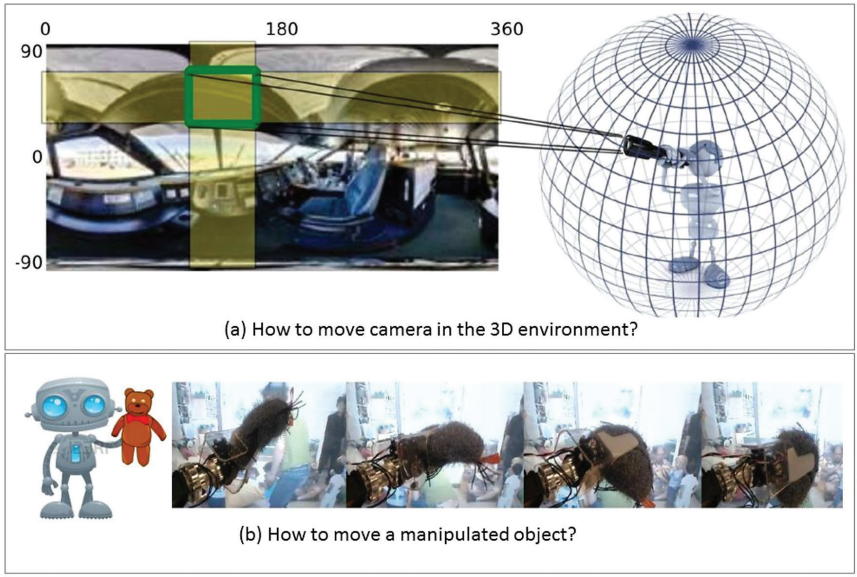

FIGURE 3 Active visual recognition requires learning how to move to reduce ambiguity in a task. A first-person vision system must learn (a) how to move its camera within the scene or (b) how to manipulate an object with respect to itself, in order to produce more accurate recognition predictions more rapidly. In (a), a robot standing in a 3D scene actively determines where to look next to categorize its environment. In (b), a robot holding an object actively decides how to rotate the object in its grasp so as to recognize it most quickly. Reprinted with permission from Jayaraman and Grauman (2016).

We show how a recurrent neural network–based system may perform end-to-end learning of motion policies suited for this "active recognition" setting. In particular, the three functions of control, per-view recognition, and evidence fusion are simultaneously addressed in a single learning objective. Results so far show that this significantly improves the capacity to recognize a scene by instructing the egocentric camera where to point next, and to recognize an object manipulated by a robot arm by determining how to turn the object in its grasp to get the sequence of most informative views (Figure 3).

EGOCENTRIC SUMMARIZATION: WHAT IS IMPORTANT IN A LONG FIRST-PERSON VIDEO?

A second major thrust of our research explores *video summarization* from the first-person perspective. Given hours of first-person video, the goal is to produce a compact storyboard or a condensed video that retains all the important people,

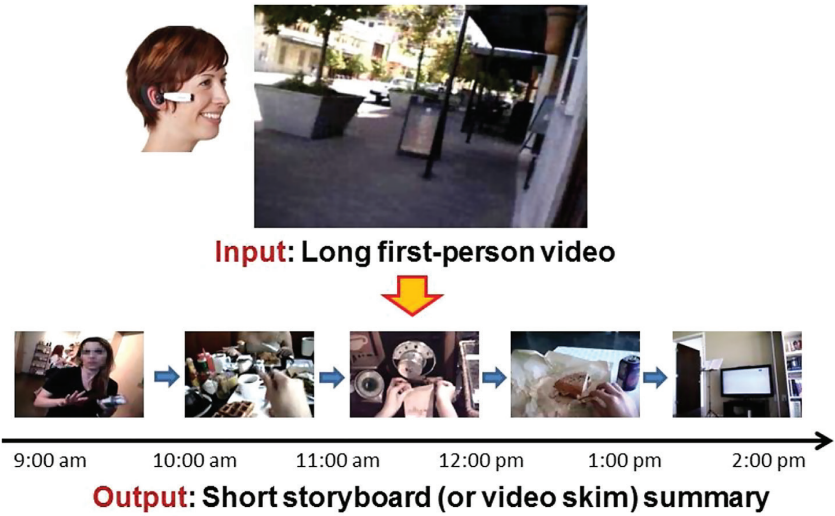

FIGURE 4 The goal in egocentric video summarization is to compress a long input video (here, depicting daily life activity) into a short human-watchable output that conveys all essential events, objects, and people to reconstruct the full story.

objects, and events from the source video (Figure 4). In other words, long video in, short video out. If the summary is done well, it can serve as a good proxy for the original in the eyes of a human viewer.

While summarization is valuable in many domains where video must be more accessible for searching and browsing, it is particularly compelling in the first-person setting because of (1) the long-running nature of video generated from an always-on egocentric camera and (2) the *storyline* embedded in the unedited video captured from a first-person perspective.

Our work is inspired by the potential application of aiding a person with memory loss, who by reviewing their visual experience in brief could improve their recall (Hodges et al. 2011). Other applications include facilitating transparency and memory for law enforcement officers wearing bodycams, or allowing a robot exploring uncharted territory to return with an executive visual summary of everything it saw.

We are developing methods to generate visual synopses from egocentric video. Leveraging cues about ego attention and interactions to infer a storyline, the proposed methods automatically detect the highlights in long source videos. Our main contributions so far entail

- learning to predict when an observed object/person is important given the context of the video (Lee and Grauman 2015),
- inferring the influence between subevents in order to produce smooth, coherent summaries (Lu and Grauman 2013),
- identifying which egocentric video frames passively captured with the wearable camera look as if they could be intentionally taken photographs (i.e., if the camera wearer were instead actively controlling a camera) (Xiong and Grauman 2015), and
- detecting temporal intervals where the camera wearer's engagement with the environment is heightened (Su and Grauman 2016).

With experiments processing dozens of hours of unconstrained video of daily life activity, we show that long first-person videos can be distilled to succinct visual storyboards that are understandable in just moments.

CONCLUSION

The first-person setting offers exciting new opportunities for large-scale visual learning. The work described above offers a starting point toward the greater goals of embodied representation learning, first-person recognition, and storylines in first-person observations.

Future directions for research in this area include expanding sensing to multiple modalities (audio, three-dimensional depth), giving an agent volition about its motions during training as well as at the time of inference, investigating the most effective means to convey a visual or visual-linguistic summary, and scaling algorithms to cope with large-scale streaming video while making such complex decisions.

REFERENCES

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 20–25, Miami Beach.

Held R, Hein A. 1963. Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology 56(5):872–876.

Hodges S, Berry E, Wood K. 2011. Sensecam: A wearable camera which stimulates and rehabilitates autobiographical memory. Memory 19(7):685–696.

Jayaraman D, Grauman K. 2015. Learning image representations tied to ego-motion. Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 13–16, Santiago. Available at www.cs.utexas.edu/~grauman/papers/jayaraman-iccv2015.pdf.

Jayaraman D, Grauman K. 2016. Look-ahead before you leap: End-to-end active recognition by forecasting the effect of motion. Proceedings of the European Conference on Computer Vision (ECCV), October 8–16, Amsterdam. Available at www.cs.utexas.edu/~grauman/papers/jayaraman-eccv2016-activerec.pdf.

Lee YJ, Grauman K. 2015. Predicting important objects for egocentric video summarization. International Journal on Computer Vision 114(1):38–55.

Lu Z, Grauman K. 2013. Story-driven summarization for egocentric video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23–28, Portland, OR. Available at www.cs.utexas.edu/~grauman/papers/lu-grauman-cvpr2013.pdf.

Su Y-C, Grauman K. 2016. Detecting engagement in egocentric video. Proceedings of the European Conference on Computer Vision (ECCV), October 8–16, Amsterdam. Available at www.cs.utexas.edu/~grauman/papers/su-eccv2016-ego.pdf.

Xiong B, Grauman K. 2015. Intentional photos from an unintentional photographer: Detecting snap points in egocentric video with a web photo prior. In: Mobile Cloud Visual Media Computing: From Interaction to Service, eds. Hua G, Hua X-S. Cham, Switzerland: Springer International Publishing.

A Quintillion Live Pixels: The Challenge of Continuously Interpreting and Organizing the World's Visual Information

KAYVON FATAHALIAN
Carnegie Mellon University

I estimate that by 2030 cameras across the world will have an aggregate sensing capacity exceeding 1 quintillion (10^{18}) pixels. These cameras—embedded in vehicles, worn on the body, and positioned throughout public and private everyday environments—will generate a worldwide visual data stream that is over eight orders of magnitude greater than YouTube's current daily rate of video ingest. *A vast majority of these images will never be observed by a human eye*—doing so would require every human on the planet to spend their life watching the equivalent of 10 high-definition video feeds! Instead, future computer systems will be tasked to automatically observe, understand, and extract value from this dense sampling of life's events.

Some applications of this emerging capability trigger clear privacy and oversight concerns, and will rightfully be subject to rigorous public debate. Many others, however, clearly have the potential for critical impact on central human challenges of the coming decades. Sophisticated image analysis, deployed at scale, will play a role in realizing efficient autonomous transportation, optimizing the daily operation of future megacities, enabling fine-scale environmental monitoring, and advancing how humans access information and interact with information technology.

The ability to develop new image understanding techniques (see Grauman in this volume), architect large-scale systems to efficiently execute these computations (the subject of my research), and deploy these systems transparently and responsibly to improve worldwide quality of life is a key engineering challenge of the coming decade.

To understand the potential impact of these quintillion pixels, let's examine

the role of image understanding in three contexts: via cameras on vehicles, on the human body, and in urban environments.

CONTINUOUS CAPTURE ON VEHICLES

It is estimated that there will be more than 2 billion cars in the world by 2030 (Sperling and Gordon 2010). Regardless of the extent to which autonomous capability is present in these vehicles, a vast majority of them will feature high-resolution image sensing. (High-resolution cameras, augmented with high-performance image processing systems, will be a low-cost and higher-information-content alternative to more expensive active sensing technologies such as Lidar.)

Image analysis systems can localize vehicles in their expected surroundings and interpret dynamic environments to predict and detect obstacles as they arise. They are thus critical to the development of vehicles that drive more safely and use roads more efficiently than human drivers. Researchers in academia and industry are racing to develop efficient image processing systems that can execute image understanding tasks simultaneously on multiple high-resolution video feeds and with low latency. Hundreds of tera-ops of processing capability—available only in top supercomputers just a decade ago—will soon be commonplace in vehicles, and computer vision algorithms are being rethought to meet the needs of these systems. High-performance analysis of vehicular video feeds will enable significant advances in transportation efficiency.

CONTINUOUS CAPTURE ON HUMANS

Although on-body cameras, such as Google Glass, have thus far failed to realize widespread social acceptance, there are compelling reasons for cameras to capture the world from the perspective of a human ("egocentric" video). For example, enabling mobile augmented reality (AR) requires systems to precisely know where a person is and what a headset wearer is looking at. (Microsoft's HoloLens headset is one example of promising recent advances in practical AR technology.) To achieve commodity, pervasive AR demands continuous, low-energy egocentric video capture and analysis.

More ambitiously, for computers to take on a more expansive role in augmenting human capabilities (e.g., the ever-present life assistant), they must "understand" much more about individuals than their present location, the contents of their inbox, and their daily calendar. For this use computers will be tasked to observe and interpret human social interactions in order to know what advice to give, and when and how to interject information.

For example, during a recent trip to Korea I found myself wishing to experience a meal at a local night market. But my inability to speak Korean and my unfamiliarity with the market's social customs made for a challenging experience in the bustling atmosphere. Imagine the utility of a system that, given a similar

view of the world as I, could not only identify the foods in front of me but also suggest how to assimilate into the crowd in front of a vendor (be assertive? attempt to form a line?), instruct me whether it was acceptable to sit in a rare open seat near a family occupying half a table (yes, it would be okay to join them), and detect and inform me of socially awkward actions I might be taking as a visitor (you are annoying the local patrons because you are violating this social norm!). These tasks illustrate how mobile computing devices will be tasked to constantly observe and interpret complex environments and complex human social interactions. Cameras located on the body, seeing the world continuously as the wearer does, are an attractive sensing modality for these tasks.

CONTINUOUS CAPTURE OF URBAN ENVIRONMENTS

It is clear that cameras will be increasingly pervasive in urban environments. It is estimated that about 280 million security cameras exist in the world today, with cities such as London, Beijing, and Chicago featuring thousands of cameras in public spaces (IHS 2015). While today's deployments are largely motivated by security concerns, the ability to sense and understand the flow of urban life both in public and private spaces provides unique opportunities to better manage modern urban challenges such as optimizing urban energy consumption, monitoring infrastructure and environmental health, and informing urban planning.

PUTTING IT ALL TOGETHER: ONE QUINTILLION PIXELS

In 2030 there will be 8.5 billion people in the world (UN 2015), 2 billion cars, and, extrapolating from recent trends (IHS 2015), at least 1.1 billion security/web cameras. Conservatively assigning one camera to each human and eight views of the road to each car, and assuming 8,000 stereo video streams per source (2×33 megapixels), there will be nearly 1 quintillion pixels across the world continuously sensing visual information.

The engineering challenge of ingesting and interpreting this information stream is immense. For example, using today's state-of-the-art machine learning methods to detect objects in this worldwide video stream would consume nearly 10^{13} watts of computing power (Nvidia 2015), even if executed on today's most efficient parallel processors. This is approximately the same amount of power used by humans around the world today (IEA 2015). Clearly, advances both in image analysis algorithms and the design of energy-efficient visual data processing platforms are needed to realize ubiquitous visual sensing.

Addressing this challenge will be a major focus of research spanning multiple areas of engineering and computer science in the coming years—machine vision, machine learning, artificial intelligence, compilation techniques, and computer architecture. Success developing these fundamental computing technologies

will provide new, valuable technology tools to tackle some of the world's most important future challenges.

REFERENCES

Grauman K. 2017. First-person computational vision. Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2016 Symposium, Washington: National Academies Press.

IEA [International Energy Agency]. 2015. 2015 Key World Energy Statistics. Paris.

IHS. 2015. Video Surveillance Camera Installed Base Report. Englewood, CO: IHS Technology.

Nvidia. 2015. GPU-Based Deep Learning Inference: A Performance and Power Analysis. Santa Clara: Nvidia Corporation.

Sperling D, Gordon D. 2010. Two Billion Cars: Driving Toward Sustainability. New York: Oxford Academic Press.

UN [United Nations]. 2015. World Population Prospects: The 2015 Revision. New York.

Extreme Engineering: Extreme Autonomy in Space, Air, Land, and Under Water

Extreme Engineering: Extreme Autonomy in Space and Air, on Land, and Under Water

DeShawn Jackson
Halliburton

Marco Pavone
Stanford University

Until now robotics systems have found application primarily in highly structured environments, for example as manipulators in an assembly line, where robotic tasks are highly repetitive and can be largely preprogrammed and the environment is carefully controlled. In the few instances where robotic systems are operated outside of the factory, they usually rely on close human supervision.

However, recent breakthroughs in decision making, perception architectures, and mechanical design, among others, are paving the way for autonomous robotic systems carrying out a wide range of tasks of unprecedented complexity—think of autonomous space vehicles, drones, self-driving cars, and unmanned underwater vehicles.

The goal of this session was to provide a representative overview of the recent algorithmic and mechanical advances that are enabling the design and deployment of robotic systems where autonomy is pushed to the extreme, resulting in exciting innovation that borders on science fiction. Specifically, the session highlighted breakthroughs at the interface of advanced decision making and bioinspired mechanical design that are enabling first-of-a-kind applications of autonomy in space (pinpoint landing of space rockets), in air (design of micro unmanned aerial vehicles), on land (high-performance legged robotic systems), and in water (autonomous underwater vehicles).

The first speaker, Lars Blackmore from Space Exploration Technologies (SpaceX), started off by discussing autonomy in space. He is the coinventor of the G-FOLD algorithm for precision landing on Mars, and his team recently completed the first precision landing of a booster stage. He discussed his work on the autonomous precision landing technology for the *Grasshopper* and F9R-Dev rockets.

Next, David Lentink from Stanford University discussed autonomous, bio-inspired micro flying robots. His innovations are revolutionizing the design of these robots, and he presented the ideas that made it possible.[1]

The session's third speaker, Sangbae Kim from MIT, addressed autonomy on land. He and other researchers at MIT have created the robotic cheetah, "the first four-legged robot to run and jump over obstacles autonomously." He explained how this robot is able to manage highly dynamic activities such as balance, energy, and impact without human interaction.[1]

Finally, Derek Paley, from the University of Maryland, looked at autonomy under water, specifically his work on motion guidance for ocean sampling by underwater vehicles.

[1] Papers not included in this volume.

Autonomous Precision Landing of Space Rockets

Lars Blackmore
SpaceX

Landing an autonomous spacecraft or rocket is very challenging, and landing one with precision close to a prescribed target even more so. Precision landing has the potential to improve exploration of the solar system and to enable rockets that can be refueled and reused like an airplane.

This paper reviews the challenges of precision landing, recent advances that have enabled precision landing on Earth for commercial reusable rockets, and what is required to extend this to landing on planets such as Mars.

BRIEF HISTORY OF AUTONOMOUS SPACE LANDINGS

In the past 50 years autonomous spacecraft have brought humans back from space, landed several rovers on the surface of Mars (Bonfiglio et al. 2011; Golombek et al. 1997; Soffen and Snyder 1976; Squyres 2005; Way et al. 2006), got a probe onto Saturn's moon Titan (Tomasko et al. 2002), landed on an asteroid (Bibring et al. 2007), and more. Because of these missions, it is now known that Mars was once warm with plenty of water and could likely have supported life, and that Titan has lakes of methane, an organic compound. Steady progress has enabled heavier payloads to be landed in more exotic locations, and recent improvements, such as advanced decelerator technologies (Tibbits and Ivanov 2015), will further expand explorers' reach in the solar system.

Although these missions have aimed for a particular location on the surface of a target planet, the precision has varied. Precision is quantified using a *landing ellipse*, the region where it is 99 percent likely that the vehicle will land. Before flight, mission planners must choose a landing site such that everywhere in the landing ellipse is safe for touchdown. Figure 1 shows that the landing ellipse for

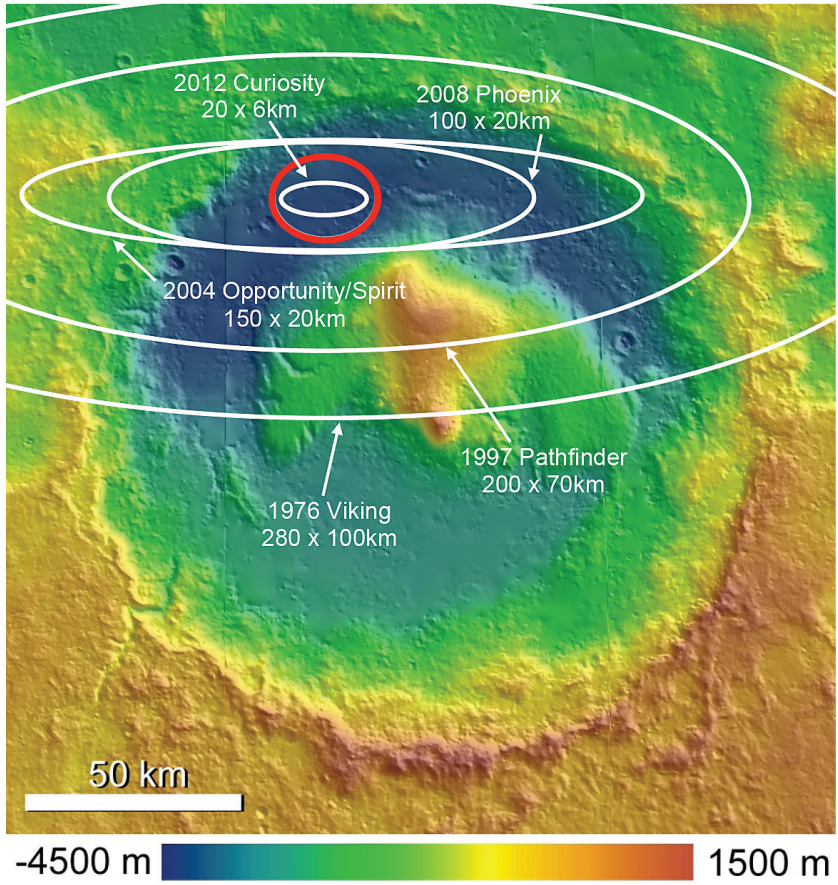

FIGURE 1 Landing ellipses for successful Mars landings to date, shown on elevation map of Gale Crater. Highlighted in the center is Curiosity's landing target, known as Aeolis Palus. Image credit: Ryan Anderson, USGS Astrogeology Science Center.

Mars missions has steadily improved, but is still measured in kilometers rather than meters.

THE NEED FOR PRECISION

When precision is measured in kilometers, missions must land in a desert (in the case of Mars) or in the ocean or on plains (in the case of Earth). If land-

ing precision could be measured in meters instead of kilometers, a world of new opportunities would open up: it would be possible to

- explore Martian caves and valleys,
- return samples from other planets,
- set up permanent outposts throughout the solar system, and
- make rockets that, after putting a payload into orbit, can be refueled and reused like an airplane, instead of being thrown away after a single flight, thus dramatically decreasing the cost of space travel.

CHALLENGES

There are some important challenges to precision landing on a planet.

Extreme Environment

A vehicle entering an atmosphere from space goes through extreme conditions.

- The majority of the entry energy is dissipated through friction with the atmosphere, resulting in extreme heating that must be dissipated; for example, the leading edge of the Apollo heatshield reached over 2500 degrees Celsius (Launius and Jenkins 2012).
- Drag causes enormous forces on the reentry vehicle; for example, SpaceX's Falcon 9 Reusable (F9R) weighs about 35 metric tons and has a peak deceleration of six times Earth gravity on reentry.
- Winds push around the reentry vehicle, with high-altitude winds at Earth regularly exceeding 100 miles per hour.
- Communication may be denied for all or part of reentry as ionized air around the spacecraft interferes with radio communications; for example, the Apollo 13 return capsule endured a 6-minute blackout.
- And finally, a spacecraft operating outside of Earth orbit is subject to high radiation, which can be fatal for electronics. This is especially true of missions operating near Jupiter, where the radiation environment is particularly intense.

Small Margin for Error

With most landings, the first attempt must be a success or the vehicle will be destroyed on impact. Moreover, additional propellant is rarely available for a second landing attempt. For large rocket engines, throttling down to a hover is technically challenging and inefficient—every second spent hovering is wasted propellant.

For F9R, the rocket has to hit zero velocity at exactly zero altitude. If it

reaches zero velocity too low, it will crash; if it reaches zero too high, it will start going back up, at which point cutting the engines and falling is the only option. This requires precise knowledge and control of vertical position and velocity.

Touchdown Challenges

A dedicated system, such as landing legs, is usually used to attenuate the loads of landing, keep the rocket safe from rocks, and prevent it from tipping over after landing. Being able to design legs that can do this as mass- and space-efficiently as possible is a challenge, as is delivering the rocket to the upright and stationary position required to avoid overloading the legs' capabilities. For the *Curiosity* rover, the SkyCrane system enabled the dual use of the rover suspension as the landing attenuation system (Prakash et al. 2008).

In addition, the landing environment may be hazardous. For the Mars Exploration rovers, the combination of rocks and high winds threatened to burst the landing airbags, so an autonomous vision and rocket system was added to detect and reduce lateral velocity (Johnson et al. 2007).

Need to Hit the Target

Achieving precision landing requires the vehicle to hit the target despite being pushed around by disturbances such as winds. For a space reentry vehicle, this is a unique problem because it is neither a ballistic missile nor an airplane. A ballistic missile tries to hit its target at high speed, so (like a bullet) it uses a high ballistic coefficient and high velocity to avoid being affected by disturbances. An airplane does get pushed around by disturbances, but its wings give it the control authority to correct for those disturbances with ease. A rocket landing vertically has neither of these advantages, making precision landing highly challenging.

RECENT ADVANCES

In the past 2 years, two commercial companies, SpaceX and Blue Origin, have sent rockets into space and landed them back on Earth within meters of their targets. Blue Origin's *New Shepard* rocket has landed several times at the company's West Texas test site. SpaceX's *Falcon 9* first stage has landed both on land at Cape Canaveral and on a floating landing platform known as the autonomous spaceport drone ship (ASDS), shown in Figure 2. Images from recent SpaceX landings are shown in Figure 3.

Central to achieving precision landing is the ability to control dispersions, which are variations in the trajectory caused by environmental uncertainty. To illustrate this, consider the example of *Falcon 9*'s first stage returning from space. To achieve precision landing, dispersions must be controlled so that, at touchdown, at least 99 percent of them fit within the designated landing zone. For F9R,

FIGURE 2 Left: SpaceX's Landing Zone 1 at Cape Canaveral. Right: The SpaceX autonomous spaceport drone ship.

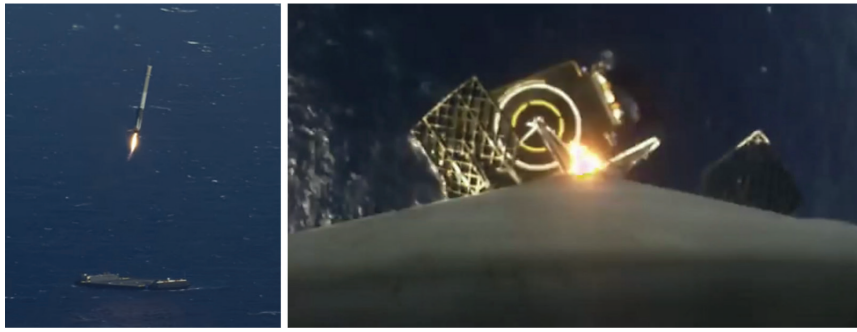

FIGURE 3 SpaceX F9R approaching the drone ship for landing.

this means achieving dispersions in the landing location of 10 meters or better for a drone ship touchdown and 30 meters or better for a landing at Cape Canaveral.

Figure 4 shows the various phases of F9R's mission. On ascent, winds push the rocket around so that dispersions grow. The first opportunity to shrink dispersions is the boostback burn, which sends the rocket shooting back toward the launch pad. During atmospheric entry, winds and atmospheric uncertainties again act to increase dispersions. The landing burn is the last opportunity to reduce the dispersions, and requires the ability to *divert*, or move sideways.

For F9R, controlling dispersions requires precision boostback burn targeting, endo-atmospheric control with fins (shown in Figure 5), and a landing burn with a divert maneuver. The latter is one of the most challenging aspects, and is also required for proposed precision landings on Mars (Wolf et al. 2011). The vehicle must compute a divert trajectory from its current location to the target, ending at

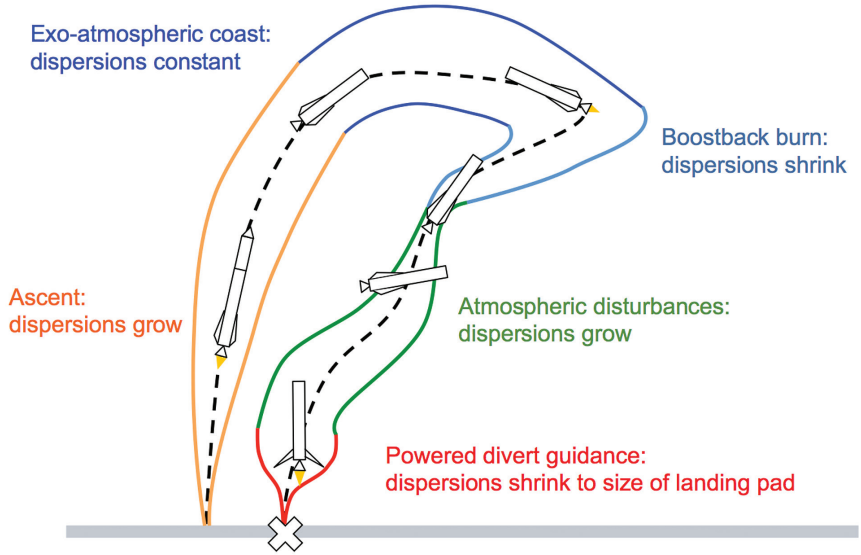

FIGURE 4 Phases of an F9R return-to-launch-site mission. The grey to black lines represent the largest possible variations in the trajectory, known as dispersions. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

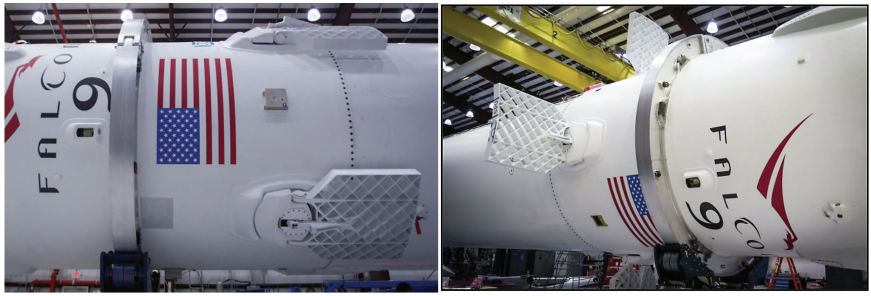

FIGURE 5 F9R's grid fins, stowed for launch (left) and deployed for entry (right).

rest and in a good orientation for landing without exceeding the capabilities of the hardware.

The computation must be done autonomously, in a fraction of a second. Failure to find a feasible solution in time will crash the spacecraft into the ground. Failure to find the optimal solution may use up the available propellant, with the same result. Finally, a hardware failure may require replanning the trajectory multiple times.

A general solution to such problems has existed in one dimension since the 1960s (Meditch 1964), but not in three dimensions. Over the past decade, research has shown how to use modern mathematical optimization techniques to solve this problem for a Mars landing, with guarantees that the best solution can be found in time (Açikmeşe and Ploen 2007; Blackmore et al. 2010).

Because Earth's atmosphere is 100 times as dense as that of Mars, aerodynamic forces become the primary concern rather than a disturbance so small that it can be neglected in the trajectory planning phase. As a result, Earth landing is a very different problem, but SpaceX and Blue Origin have shown that this too can be solved. SpaceX uses CVXGEN (Mattingley and Boyd 2012) to generate customized flight code, which enables very high-speed onboard convex optimization.

NEXT STEPS

Although high-precision landings from space have happened on Earth, challenges stand in the way of transferring this technology to landing on other bodies in the solar system.

One problem is navigation: precision landing requires that the rocket know precisely where it is and how fast it's moving. While GPS is a great asset for Earth landing, everywhere else in the universe is a GPS-denied environment. Almost all planetary missions have relied on Earth-based navigation: enormous radio antennas track the vehicle, compute its position and velocity, and uplink that information to the vehicle's flight computer. This is sufficient for landings that only need to be precise to many kilometers, but not for landings that need to be precise to many meters.

Analogous to driving while looking in the rearview mirror, Earth-based tracking gets less and less accurate at greater distances from the starting point. Instead, the focus needs to be on the destination planet in order to be able to land precisely on it. *Deep Impact* is an example of a mission that used its target to navigate (Henderson and Blume 2015), but (as its name implies) it was an impactor mission, not a landing.

Recent research has achieved navigation accuracy on the order of tens of meters (Johnson et al. 2015; Wolf et al. 2011) using terrain relative navigation, where the lander images the surface of the planet during landing and matches features with an onboard map to determine its location. This can be tested on Earth, at least in part, without the need to perform the entire reentry from space.

Several companies have used experimental vehicles, some of which are shown in Figure 6, to demonstrate powered descent technology with low-altitude hops. Using these vehicles, terrain relative navigation has been tested on Earth (Johnson et al. 2015), and a demonstration on Mars is being considered for the Mars 2020 rover mission. If this is successful, combining terrain relative naviga-

FIGURE 6 Various experimental vertical takeoff and landing testbeds. Clockwise from top left: NASA's *Morpheus* (left) and *Mighty Eagle* (right), Masten Aerospace's *Xoie*, SpaceX's *Grasshopper*, McDonnell Douglas' *DC-X*, and Armadillo Aerospace's *Mod*. Image credits: NASA/Dimitri Gerondidakis (*Morpheus*); NASA/MSFC/Todd Freestone (*Mighty Eagle*); Ian Kluft (*Xoie*); NASA (*DC-X*); Armadillo Aerospace/Matthew C. Ross (*Mod*).

tion with demonstrated precision guidance and control could finally make precision landings on Mars, Europa, and other bodies in this solar system a reality.

REFERENCES

Açikmeşe B, Ploen SR. 2007. Convex programming approach to powered descent guidance for Mars landing. Journal of Guidance, Control, and Dynamics 30(5):1353–1366.

Bibring J-P, Rosenbauer H, Boehnhardt H, Ulamec S, Biele J, Espinasse S, Feuerbacher B, Gaudon P, Hemmerich P, Kletzkine P, and 9 others. 2007. The Rosetta Lander ("Philae") investigations. Space Science Reviews 128(1–4):205–220.

Blackmore L, Açikmeşe B, Scharf DP. 2010. Minimum landing error powered descent guidance for Mars landing using convex optimization. Journal of Guidance, Control, and Dynamics 33(4):1161–1171.

Bonfiglio EP, Adams D, Craig L, Spencer D, Arvidson R, Heet T. 2011. Landing-site dispersion analysis and statistical assessment for the Mars *Phoenix* lander. Journal of Spacecraft and Rockets 48(5):784–797.

Golombek M, Cook RA, Economou T, Folkner WM, Haldemann AFC, Kallemeyn PH, Knudsen JM, Manning RM, Moore HJ, Parker TJ, and 4 others. 1997. Overview of the Mars Pathfinder mission and assessment of landing site predictions. Science 278(5344):1743–1748.

Henderson M, Blume W. 2015. *Deep Impact*: A review of the world's pioneering hypervelocity impact mission. Procedia Engineering 103(2015):165–172.

Johnson A, Willson R, Cheng Y, Goguen J, Leger C, San Martin M, Matthies L. 2007. Design through operation of an image-based velocity estimation system for Mars landing. International Journal of Computer Vision 74:319–341.

Johnson AE, Cheng Y, Montgomery JF, Trawny N, Tweddle B, Zheng JX. 2015. Real-time terrain relative navigation test results from a relevant environment for Mars landing. AIAA Guidance, Navigation, and Control Conference (AIAA 2015-0851), January 5–9, Kissimmee, FL.

Launius RD, Jenkins DR. 2012. Coming Home: Reentry and Recovery from Space. NASA Aeronautics Book Series. Washington.

Mattingley J, Boyd S. 2012. CVXGEN: A code generator for embedded convex optimization. Optimization and Engineering 13(1):1–27.

Meditch JS. 1964. On the problem of optimal thrust programming for a lunar soft landing. IEEE Transactions on Automatic Control 9(4):477–484.

Prakash R, Burkhart D, Chen A, Comeaux K, Guernsey C, Kipp D, Lorenzoni L, Mendeck G, Powell R, Rivellini T, and 3 others. 2008. Mars science laboratory entry, descent, and landing system overview. Proceedings of the IEEE Aerospace Conference, March 1–8, Big Sky, MT.

Soffen GA, Snyder CW. 1976. First Viking mission to Mars. Science 193:759–766.

Squyres SW. 2005. Roving Mars: Spirit, Opportunity, and the Exploration of the Red Planet. New York: Hyperion.

Tomasko MG, Buchhauser D, Bushroe M, Dafoe LE, Doose LR, Eibl A, Fellows C, Farlane EM, Prout GM, Pringle MJ. 2002. The Descent Imager/Spectral Radiometer (DISR) experiment on the Huygens entry probe of Titan. Space Science Reviews 104(1/2):467–549.

Tibbits B, Ivanov M. 2015. Low density supersonic decelerator flight dynamics test-1 flight design and targeting. 23rd AIAA Aerodynamic Decelerator Systems Technology Conference (AIAA 2015-2152), March 30–April 2, Daytona Beach.

Way D, Powell R, Chen A, Steltzner A, San Martin M, Burkhart D, Mendeck G. 2006. Mars science laboratory entry, descent, and landing system. 2006 IEEE Aerospace Conference, March 4–11, Big Sky, MT.

Wolf AA, Açikmeşe B, Cheng Y, Casoliva J, Carson JM, Ivanov MC. 2011. Toward improved landing precision on Mars. IEEE Aerospace Conference, March 5–12, Big Sky, MT.

Autonomy Under Water: Ocean Sampling by Autonomous Underwater Vehicles

Derek A. Paley
University of Maryland

The ocean is large and opaque to electromagnetic waves used for communication and navigation, making autonomy under water essential because of intermittent, low-bandwidth data transmission and inaccurate underwater positioning. Moreover, the vastness of the ocean coastlines and basins renders traditional sampling time-consuming and expensive; one typically must make do with sparse observations, which leads to the question of where to take those measurements. Further complicating the problem, circulating ocean processes pertinent to national defense and climate change span small to large space and time scales, necessitating multiple sampling platforms or vehicles, ideally with long endurance.

Many effective vehicle designs are not propeller-driven but rather buoyancy-driven; becoming heavy or light relative to the surrounding seawater generates vertical motion suitable for collecting profiles of hydrographic properties such as salinity, density, and temperature. Buoyancy-driven vehicles drift with the currents unless they have wings, in which case their vertical motion is converted to horizontal motion via lift, like an air glider. The capacity to maneuver relative to the flow field gives rise to challenges in cooperative control and adaptive sampling with the following recursive property: vehicles collect measurements of the ocean currents in order to estimate it; the estimate is used to guide the collection of subsequent measurements along sampling trajectories subject to currents that may be as large as the platform's through-water speed.

Approaches to adaptive sampling of continuous environmental processes are distinguished by the characterization of the estimated process as statistical or dynamical. A statistical characterization involves spatial and temporal decorrelation scales, which for a nonstationary process may vary in space and time.

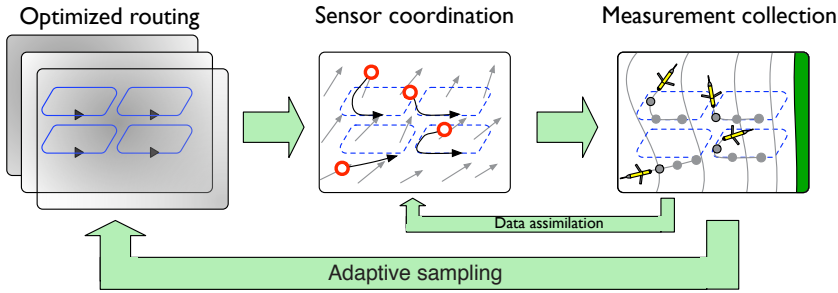

FIGURE 1 Dynamic, data-driven sampling uses (left) information-based metrics to optimize sensor routes, (middle) multivehicle control to stabilize the desired trajectories, and (right) nonlinear filtering to assimilate data; adaptive sampling refers to the reoptimization of sensor routes that occurs after data assimilation.

Presuming these scales are known, the challenge is to distribute measurements proportionally to the local variability—highly variable regions require higher measurement density—to minimize the so-called mapping error. If the scales are unknown, the challenge is twofold: they must be estimated while concurrently using estimated scales for mapping.

A dynamical characterization of the estimated process replaces the decorrelation statistics with the differential equations that govern the evolution in time of the process parameters. Dynamical descriptions permit the application of tools from systems theory, including the concept of observability, which measures the sensitivity of a system's outputs to perturbations of the system's states about a nominal value. The selection of where to collect sensor observations (i.e., where to route the sensor platforms) is thus formulated as the problem of maximizing the observability of a nonlinear dynamical system. Because observability is traditionally a forward-looking metric, it is augmented with estimation uncertainty in order to account for locations of prior observations. Figure 1 depicts the adaptive-sampling feedback loop.

DISTRIBUTED ESTIMATION OF SPATIOTEMPORAL FIELDS

Observability-based optimization in path planning (Yu et al. 2011) and data assimilation (Krener and Ide 2009) typically uses either a low-dimensional parameterized model or an empirical data-based representation of the unknown process; however, problems arise when neither a suitable model nor sufficient data are available. The novelty of the approach described here lies in the use of the observability of a low-dimensional model of an environmental vector field and a data assimilation filter to guide the observability analysis via metrics from Bayesian experimental design. Observability- and information-based optimization

of sampling trajectories yields a reliable and predictable capability for intelligent, mobile sensors. A dynamic, data-driven Bayesian nonlinear filter exploits noisy, low-fidelity, and nonlinear measurements collected in a distributed manner by combining observations from individual or multiple sampling platforms.

Although methods exist for the optimization of sampling trajectories using distributed parameter estimation (Demetriou 2010), optimal interpolation (Leonard et al. 2007), and heuristic approaches (Smith et al. 2010), an open question is how to rigorously characterize the variability of information content in an unknown spatiotemporal process and how to target observations in information-rich regions. The merit of the approach described here lies in the design of a statistical framework based on spatiotemporal estimation of nonstationary processes in meteorology (Karspeck et al. 2012) and geostatistics (Higdon et al. 1999). The framework extends my previous work in this area into multiple dimensions and builds a nonstationary statistical representation of a random process while simultaneously optimizing the sampling trajectories.

In oceanography, autonomous underwater vehicles are used as mobile sensors for adaptive sampling. Indeed, the concept of optimal experimental design was first applied to oceanographic sampling in the 1980s (Barth and Wunsch 1990). The optimization of sensor placement and data collection has applications in fighting wildfires, finding perturbation sources in power networks, and collecting spatial data for geostatistics. Perhaps it is not surprising, given the range of these applications, that there are a variety of approaches advocated in the literature, including adaptation based on maximum a posteriori estimation, stochastic deployment policies, information-based methods, learning and artificial intelligence, deterministic methods with heuristic metrics, bioinspired source localization and gradient climbing, and nonparametric Bayesian models.

The results described here differ from prior work on adaptive sampling of dynamical systems and random processes in the novel application of nonlinear observability and control coupled with recursive Bayesian filtering to optimize sensor routing for environmental sampling. One approach to adaptive sampling in the ocean uses observability: a measure of how well the state variables of a control system can be determined by measurements of its outputs.

Observability of a linear system (Hespanha 2009) is characterized using the Kalman rank condition, which is a special case of the observability rank condition of a general, nonlinear system (Hermann and Krener 1977). (A nonlinear system is called *observable* if two states are indistinguishable only when the states are identical.) Observability in data assimilation refers to the ability to determine the parameters of an unknown process from a history of observations.

Although standard observability tests give a binary answer (i.e., the system is observable or not), the degree of observability may be computed from the singular values of the observability gramian (Krener and Ide 2009), a Hermitian matrix containing inner products of the system's outputs under systematic perturbations of the system's parameters about a nominal value. Computation of the empiri-

cal observability gramian requires only the ability to simulate the system and is therefore particularly attractive for numerical optimization.

A second approach to adaptive sampling is based on classical estimation theory (Liebelt 1967): optimal statistical interpolation of sensor observations to produce a stochastic estimate of an unknown random process, formerly known in meteorology and oceanography as objective analysis (Bretherton et al. 1976). Optimal interpolation also yields a measure of the uncertainty or error in the estimate, which can be used as a measure of estimator performance or skill (Leonard et al. 2007). It is common to compute estimation error under the assumption of stationarity of the spatial and temporal variability of the unknown process, although these assumptions may not be borne out in applications of interest. A stochastic process whose variability changes when shifted in time or space is called *nonstationary*, and methods exist to parameterize nonstationary processes in oceanography and geostatistics. Indeed, nonstationary-based strategies have been applied to mobile sensor networks, though not based on a principled control design.

DATA-DRIVEN ADAPTIVE SAMPLING: MEASURES OF OBSERVABILITY

The observability of a nonlinear system may be difficult to determine analytically, because it requires tools from differential geometry (Hermann and Krener 1977). If the dynamical model of interest is solved numerically, numerical techniques can be used to calculate the empirical observability gramian (Krener and Ide 2009). This gramian does not require linearization, which may fail to adequately model the input-output relationship of the nonlinear system over a wide range of operating conditions, but merely the ability to simulate the system. Indeed, it maps the input-output behavior of a nonlinear system more accurately than the observability gramian produced by linearization of the nonlinear system.

The empirical observability gramian is a square matrix whose dimension matches the size of the state vector and whose (i,j)th entry represents the sensitivity of the output to infinitesimal perturbations about their nominal value of the corresponding i and j states or unknown parameters. The observability of a nonlinear system is measured by calculating the unobservability index ν of the empirical observability gramian. This index is used to score candidate trajectories for their anticipated information gain. Figure 2 depicts an observability-based feedback loop, including a recursive Bayesian filter to assimilate noisy measurements from the observing platforms.

DATA-DRIVEN ADAPTIVE SAMPLING: MAPPING ERROR

A spatiotemporal field is statistically described by its mean and the covariance function between any two points i and j. A covariance function is a positive-

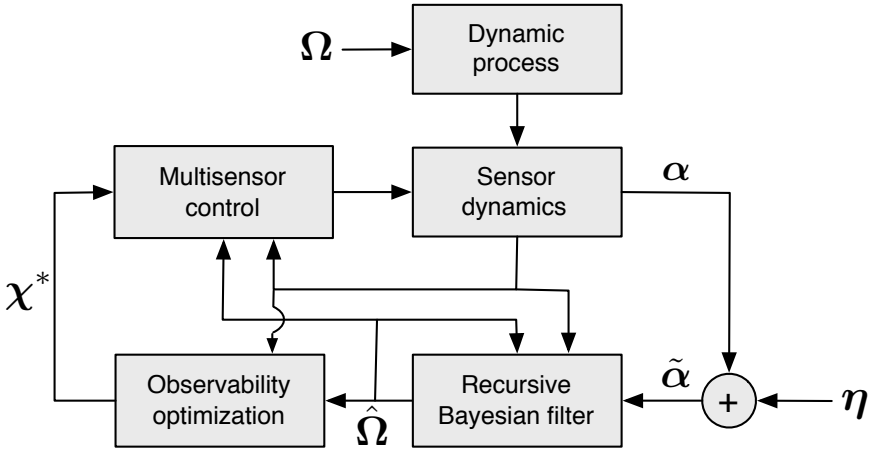

FIGURE 2 Schematic diagram of an observability-based sampling algorithm. A recursive Bayesian filter provides parameter estimates $\overline{\Omega}$ from noisy measurements $\tilde{a}$ corrupted by noise η. The estimated parameters are used to calculate observability-optimizing control parameters χ^* that characterize the multisensor sampling formation.

definite function that describes the variability of the field between the *i*th and *j*th locations (Bennett 2005). A field is stationary if its covariance function depends only on the relative position of *i* and *j*, and it is nonstationary if it depends on *i* and *j* independently.

There are a number of choices for the form of a nonstationary covariance function, e.g., Matern, rational quadratic, Ornstein-Uhlenbeck, and squared-exponential forms (see Higdon et al. 1999). A statistics-based sampling strategy requires a covariance function that is a product of spatial- and temporal-covariance functions, such as a nonstationary squared exponential covariance function. In this case, the square roots of the diagonal elements are the spatial and temporal decorrelation scales of the field. (The decorrelation scales are the spatial and temporal separations at which the covariance function evaluates to $1/e$.) For a stationary field the decorrelation scales are constant, whereas for a nonstationary field they may vary in space and time. The covariance function is used to derive a coordinate transformation that clusters measurements in space-time regions with short decorrelation scales and spreads measurements elsewhere, where the measurement demand is lower.

Statistics-based sensor routing seeks to provide optimal coverage of an estimated spatiotemporal field. The coverage is deemed optimal when the measurement density in space and time is proportional to the variability of the field. To determine when measurements are redundant, consider the footprint of a measurement, defined as the volume in space and time in an ellipsoid centered at the

measurement location with principal axes equal to the decorrelation scales of the field. The goal is to design the vehicle trajectories so that the swaths created by the set of all measurement footprints cover the entire field with minimal overlaps or gaps, even when the decorrelation scales of the field vary.

Optimal interpolation is used to determine the mapping error (Bennett 2005), which is the diagonal of the error covariance matrix. The average (resp. maximum) mapping error is computed by averaging (resp. finding the maximum of) all the elements of the mapping error. Because the vehicles sample uniformly in time, the mapping error is minimized in a stationary field by traveling at maximum speed to place as many measurements as possible in the domain (Sydney and Paley 2014). Figure 3 depicts the mapping error for a stationary field with a correlation scale estimated by a Bayesian filter.

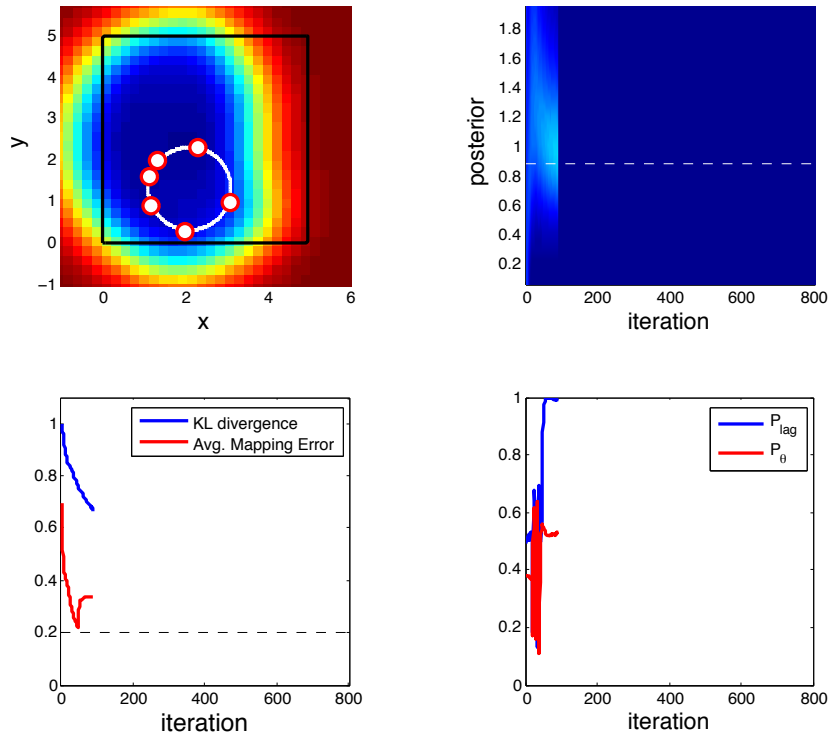

FIGURE 3 (left) Mapping error of a 2D stationary spatiotemporal field, where x and y represent two spatial dimensions; (right) dynamic estimation of the decorrelation scale. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

REFERENCES

Barth N, Wunsch C. 1990. Oceanographic experiment design by simulated annealing. Journal of Physical Oceanography 20(9):1249–1263.

Bennett AF. 2005. Inverse Modeling of the Ocean and Atmosphere. Cambridge, UK: Cambridge University Press.

Bretherton FP, Davis RE, Fandry CB. 1976. A technique for objective analysis and design of oceanographic experiments applied to MODE-73. Deep-Sea Research 23(7):559–582.

Demetriou MA. 2010. Guidance of mobile actuator-plus-sensor networks for improved control and estimation of distributed parameter systems. IEEE Transactions on Automatic Control 55(7):1570–1584.

Hermann R, Krener AJ. 1977. Nonlinear controllability and observability. IEEE Transactions on Automatic Control 22(5):728–740.

Hespanha J. 2009. Linear Systems Theory. Princeton, NJ: Princeton University Press.

Higdon D, Swall J, Kern J. 1999. Non-stationary spatial modeling. Bayesian Statistics 6(1):761–768.

Karspeck AR, Kaplan A, Sain SR. 2012. Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in North Atlantic sea-surface temperatures for 1850–2008. Quarterly Journal of the Royal Meteorological Society 138(662):234–248.

Krener AJ, Ide K. 2009. Measures of unobservability. Proceedings of the IEEE Conference on Decision and Control, pp. 6401–6406, Shanghai, December 16–18.

Leonard NE, Paley DA, Lekien F, Sepulchre R, Fratantoni DM, Davis RE. 2007. Collective motion, sensor networks and ocean sampling. Proceedings of the IEEE 95(1):48–74.

Liebelt PB. 1967. An Introduction to Optimal Estimation. Reading, MA: Addison-Wesley.

Smith SL, Schwager M, Rus D. 2010. Persistent robotic tasks: Monitoring and sweeping in changing environments. IEEE Transactions on Robotics 28(2):410–426.

Sydney N, Paley DA. 2014. Multivehicle coverage control for nonstationary spatiotemporal fields. Automatica 50(5):1381–1390.

Yu H, Sharma R, Beard RW, Taylor CN. 2011. Observability-based local path planning and collision avoidance for micro air vehicles using bearing-only measurements. Proceedings of the American Control Conference, pp. 4649–4654, June 29–July 1, San Francisco.

Water Desalination and Purification

Water Desalination and Purification

Amy Childress
University of Southern California

Abhishek Roy
The Dow Chemical Company

Securing a reliable supply of water among growing human populations, changing climate, and increasing urbanization is a global challenge. Water-stressed regions are exploring alternative sources to augment their freshwater supplies. This session focuses on membrane separation processes to desalinate and purify a range of source waters. Innovations in materials, developments in new processes, and synthesis of novel systems are emphasized for applications spanning desalination, wastewater reclamation, and treatment of industrial streams with complex solution chemistries.

Reverse osmosis (RO) desalination is currently the most efficient and widely adopted commercial desalination technology; however, it requires a great deal of energy to create the high pressures necessary to overcome the osmotic pressure of saline waters and there are often significant issues with disposal of brine resulting from the process. Technological advances are needed to improve the energy efficiency, contaminant removal, and environmental impacts of the processes. Current focus is on improving sustainability in conventional applications such as sea- and brackish-water desalination and exploring emerging opportunities in municipal and industrial wastewater desalination markets.

Membranes with high flux, high rejection, and low tendency for fouling are most desired for use in emerging and conventional treatment processes that provide consistently high process performance, require few chemicals, and produce little waste. The goal is to develop highly efficient, reliable, and durable treatment systems that can be scaled for use in distributed or centralized applications. Centralized systems capitalize on use of existing infrastructure; smaller-scale, distributed systems may offer better opportunities for coupling treatment with

alternative energy sources. Sustainable system performance is key for efficiency and economics as well as adherence to health and regulatory standards.

The session provided a forum for the audience to discuss and identify collaborative opportunities in four critical areas of water desalination and purification: new materials development, analytical characterization techniques, emerging desalination technologies, and innovative system design and operation. It began with a talk by Manish Kumar (Pennsylvania State University), a high-level overview of RO technology, applications, and recent membrane chemistry innovations.

Chris Stafford (National Institute of Standards and Technology) delved into state-of-the-art polyamide membrane chemistries, emphasizing the importance of advanced membrane characterization techniques to drive breakthrough innovations.

Baoxia Mi (University of California, Berkeley) introduced emerging desalination treatment technologies and highlighted new materials being developed to further advance these technologies.

Kevin Alexander (Hazen and Sawyer) wrapped up the session with a techno-economic assessment of high-recovery treatment from impaired waters, including applications with challenging solution chemistries and efforts to achieve zero liquid discharge.

Water Desalination: History, Advances, and Challenges

Manish Kumar, Tyler Culp, and Yuexiao Shen
Pennsylvania State University

Desalination is the removal of salt and contaminants from water. It involves a broad range of technologies that yield access to marginal sources of water such as seawater, brackish ground- and surface water, and wastewater. Given the reduction in access to fresh water in recent decades and the uncertainty in availability effected by climate change, desalination is critical for ensuring the future of humanity.

This paper describes advances toward more sustainable desalination and exciting directions that could make this technology more accessible, energy efficient, and versatile. It reviews the emergence of membranes as the preferred technology for desalination, recent advances, challenges to its sustainable implementation, and needs for further research.

INTRODUCTION

Desalination represents a promise of near unlimited water supply and is an attractive potential solution to the age-old conundrum of seawater abundance and practical inaccessibility for potable use. It now encompasses the removal of both salts and dissolved contaminants from various sources such as seawater, brackish surface and groundwater, and industrial and municipal wastewaters.

The primary descriptor of importance for desalination processes is the amount of dissolved solids (primarily inorganic salts) represented by the total dissolved solids (TDS; the solids left over after water is evaporated from particle-free water). Table 1 lists the typical range of TDS levels in waters subjected to desalination-based water treatment processes (Australian NWC 2008).

In addition to being a measure of usability (such as for consumption), TDS

TABLE 1 Typical Water Sources for Desalination and Their Total Dissolved Solids (TDS) Ranges as Well as the Calculated Minimum Energy for Separation per Unit Volume (specific energy consumption).

Water Source*	Total Dissolved Solids (mg/L)	Minimum Energy for Separation (kwh/m^3)**
Seawater	15,000–50,000	0.67
Brackish water	1,500–15,000	0.17
River water	500–3,000	0.04
Pure water	< 500	< 0.01
Wastewater (untreated domestic)	250–1,000	0.01
Wastewater (treated domestic)	500–700	0.01

* Data from Australian NWC (2008).
** Calculated based on average TDS of the range.

levels determine the bounds for the minimum energy needed to remove these solutes from water (or to move water away from these solutes). Just as energy is released when a solute is dissolved in a compatible solvent, energy is needed to separate the solute from the solvent and is dependent on the concentration of the solute. Table 1 shows that higher-salinity water (such as seawater) requires larger amounts of energy for desalination, whereas water from low-salinity streams (such as those from wastewater reuse) could be much lower.

The growing pressure on limited freshwater sources has focused the world's attention on seawater and the recovery of water from marginal sources such as brackish ground- and surface water as well as recycled wastewater. It has also raised awareness and catalyzed the implementation of wastewater reuse, where wastewater is treated to a high quality and in some cases used for direct or indirect potable reuse. Desalination is thus a critical technology for humanity to allow for sustainable development.

BACKGROUND AND HISTORY

Desalination has a long history in both mythology and practice. An early and illustrative reference appears in the Bible (Exodus 15:22–26) and is widely considered to be about desalination.

> When they came to Marah, they could not drink the water of Marah because it was bitter; therefore it was named Marah. And the people grumbled against Moses, saying, "What shall we drink?" And he cried to the LORD, and the LORD showed him a log, and he threw it into the water, and the water became sweet.

Distillation-Based Technologies

Early scientific descriptions of desalination centered around the application of distillation. In his *Meteorologica*, Aristotle wrote that "Salt water when it turns into vapour becomes sweet and the vapour does not form salt water again when it condenses" (Forbes 1948, p. 383). This is the definition of distillation, a process used to create fresh water from seawater at larger scales starting in the 1930s (NRC 2008). Distillation-based technologies remained a major approach to water desalination until the development of membranes.

The most common distillation-based desalination methods are thermally driven technologies, including multistage flash distillation, multiple-effect distillation, and mechanical vapor compression processes. In these processes water is evaporated by the addition of heat and in many cases assisted by the use of vacuum. The evaporated water is then condensed to recover desalinated water. Several large plants, primarily in the Middle East, have used thermal distillation since the 1930s (NRC 2008).

But thermal desalination has very high energy consumption and is increasingly being replaced by the use of membranes, specifically reverse osmosis (RO) membranes. Figure 1 shows the energy consumption per unit volume of water

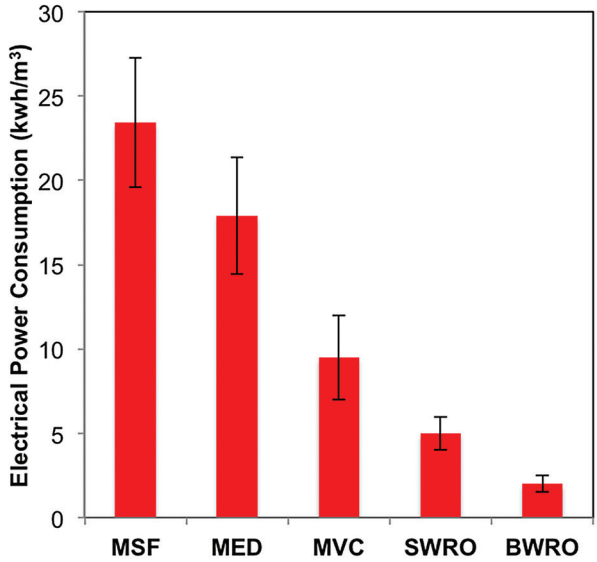

FIGURE 1 Typical equivalent (specific) electrical power consumption for thermal and membrane distillation strategies (based on data from Al-Karaghouli and Kazmerski 2013). BWRO = brackish water reverse osmosis; MED = multiple-effect distillation; MSF = multistage flash distillation; MVC = mechanical vapor compression; SWRO = seawater reverse osmosis.

for several commonly used water desalination techniques (Al-Karaghouli and Kazmerski 2013). As is evident from this figure, RO is a substantially more energy-efficient technology for water desalination.

Emergence of Membrane Technology

Membrane technologies arose as a result of a breakthrough in the use of polymer films for separating salt from water in the late 1950s/early 1960s. A brief history of the development of RO membranes is shown in Figure 2, based on Baker (2004).

Reid and Breton (1959) first demonstrated the possibility of desalination using polymeric cellulose films, and thus the first polymeric RO membranes were created. Loeb and Sourirajan (1963) then showed that an asymmetric cellulose acetate membrane can be used for desalination. The permeabilities of these early membranes were low, and RO membranes were considered a novelty separation technique rather than a solution to desalination.

An innovation in the packaging of large membrane areas into small volumes was the development of the spiral wound module (Figure 3) by General Atomics in 1963. The spiral wound configuration is now common in RO applications (Cadotte 1981; Westmoreland 1968). In this module, "leaves" of membranes,

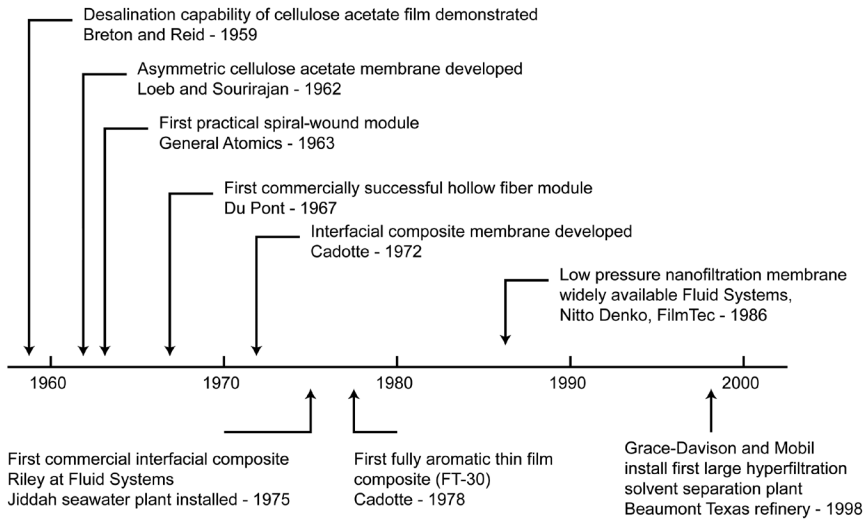

FIGURE 2 Brief timeline of the development of reverse osmosis membranes. Reproduced with permission from Baker (2004). © 2004 Wiley.

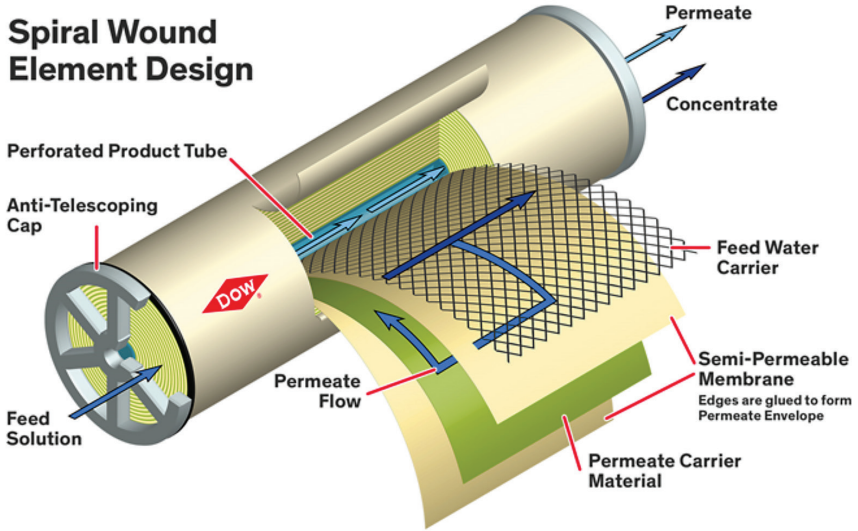

FIGURE 3 Typical spiral wound module design (used with permission from Dow Chemical).

with feed and permeate spacers, are connected to a perforated permeate tube and rolled up in a "jelly roll" configuration. Hollow-fiber modules containing thin fibers were developed a few years later by DuPont, but this configuration is less commonly used for RO.

A major advance in membrane chemistry that has made possible the application of RO membranes is the development of the thin film composite (TFC) architecture. Previously, membranes were either several-micron-thick polymer layers with a uniform architecture or similar-size polymer layers with an "asymmetric" structure with a nonporous salt-rejecting top surface opening up to a more porous support.

Cadotte (1981) patented the design for the three-layer TFC membrane that is now the industry standard. It provides high permeability while maintaining selectivity for water (vs. salt or other solutes). His major innovation was to make the crosslinked "active layer" of the membrane of nanoscale thickness and support it on a microporous membrane (Figure 4). A 20–200 nm thin crosslinked polyamide layer is supported on (or indeed grown from) a microporous polysulfone layer that is in turn supported on a polyester fabric.

The most common chemistry for modern RO membranes is interfacial polymerization, another major advance in RO membrane manufacturing. The procedure, described in Figure 5, has been the standard for making RO membranes for the past 5 decades.

The energy consumption of RO technology has dramatically declined

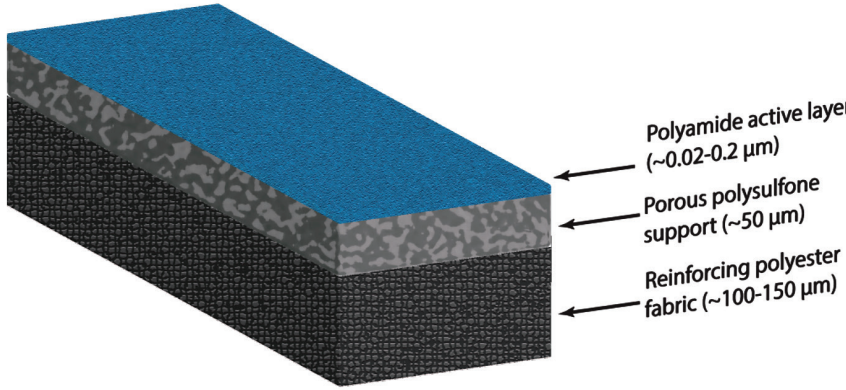

FIGURE 4 Architecture of a thin film composite (TFC) reverse osmosis (RO) membrane. A crosslinked polyamide nonporous active layer is supported on a microporous polysulfone membrane cast on a polyester fabric.

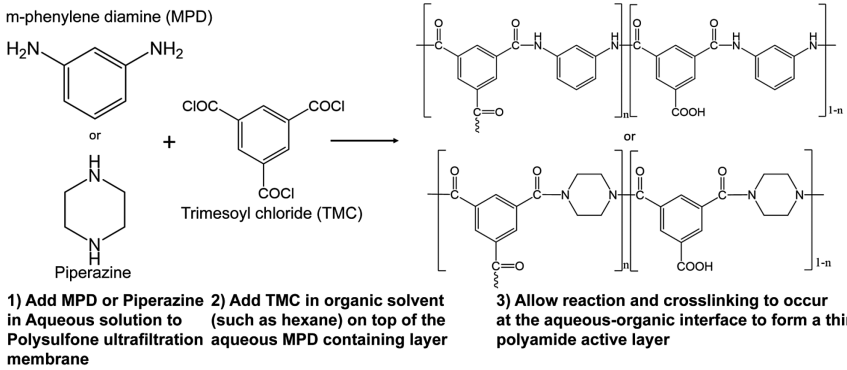

FIGURE 5 The reaction scheme and procedure most commonly used for synthesizing thin film composite (TFC) reverse osmosis (RO) and nanofiltration membranes (NF). RO membranes are typically synthesized using the MPD aqueous monomer while NF membranes are more commonly synthesized using the piperazine monomer. TMC is used for both types of membranes.

(Figure 6, based on data from Gude 2011 and Elimelech and Phillip 2011) thanks to improvements in formulation, manufacturing procedures, and processes, such as energy recovery from pressurized brine. These advances rapidly enhanced sustainability and exponentially increased the implementation of these membranes for seawater and brackish water desalination as well as wastewater reuse.

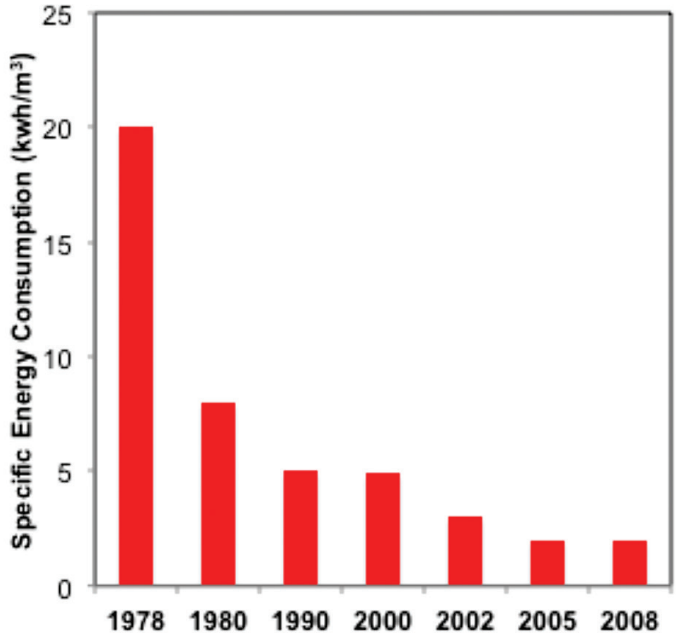

FIGURE 6 Decline in specific energy consumption of reverse osmosis membranes, 1978–2008.

For some cases, such as seawater reverse osmosis, it is argued that current membranes have reached very close to the thermodynamic limit of ~1 kWh/m^3 and that further improvement in materials may not yield additional energy sustainability (Elimelech and Phillip 2011). On the other hand, advances in permeability and selectivity can still yield major gains in brackish water treatment and wastewater reuse.

Ultrapermeable membranes with very high salt rejection appropriate for reverse osmosis may substantially reduce the necessary energy (~45 percent) or plant infrastructure (pressure vessels, up to 65 percent) in low-salinity sources (Cohen-Tanugi et al. 2014) such as brackish water desalination and water reuse. The energy advantage is significantly lower for high-salinity seawater applications (15 percent less energy) but the plant size can be reduced by 44 percent (Cohen-Tanugi et al. 2014).

A focus on increasing selectivity rather than simply increasing membrane permeability has been proposed in recent work as a sustainable approach to improve membrane materials (Werber et al. 2016a).

ADVANCES IN RO DESALINATION

Recent advances in desalination membranes promise a path to higher sustainability. Some of these advances are described below.

Channel-Based Membranes as an Alternative to Solution-Diffusion RO Membranes

RO membranes currently rely on the solution-diffusion mechanism to separate solutes from water, a transport method in which components of the solution first dissolve into the membrane matrix and then diffuse across the membrane by "jumping" between transiently connected pores. In contrast, biological membranes conduct efficient and selective channel-based transport, in which water or selected solutes are transported "straight through" protein channels (membrane proteins, MPs). MP channels are approximately 4 nm in length in comparison to the tortuous unconnected pores in the 20–200 nm thick RO membrane active layers.

Attention has recently been focused on water channel proteins called aquaporins (AQPs) and their synthetic analogs, carbon nanotubes (CNTs). AQPs selectively transport water across cell membranes in many forms of life (including in humans) (Agre 2004).

Both AQPs and CNTs efficiently transport water at the rate of several billions of molecules per second. They consist of narrow pores lined with hydrophobic surfaces, resulting in single-file water transport (de Groot and Grubmuller 2001; Hinds 2007). While CNTs cannot be made at dimensions that are substantially less than 10 Å in diameter and thus cannot reject salt (hydrated sodium and chloride ions are about 7.2 and 6.6 Å in diameter respectively; Israelachvili 2011), AQPs are highly water selective due to their small pore size (~3 Å) and the presence of amino acid residues that reject charged ions (Agre et al. 2002). The exceptional permeability and selectivity of AQPs has led to research on their incorporation in water purification membranes (Shen et al. 2014), and AQP-based biomimetic membranes were proposed in the mid- to late 2000s in several patents and papers.

There have been many advances since, including methods to incorporate AQPs in stable lipids and lipid-like block copolymers, their packing at high density into membranes, the integration of such layers into various membrane architectures, and finally the development of a scalable membrane where AQPs are inserted into the active layer of RO membranes (Zhao et al. 2012). The latter has resulted in commercially available membranes at small scale, but they face significant challenges to scaleup because of concerns about stability and cost.

Another advance inspired by biological channels and arguably more scalable is the development of artificial water channels and proposals to develop membranes around them (Barboiu 2012). These bioinspired channels are made synthetically using organic synthesis but have until recently been a less studied

area with only a few architectures reported (Shen et al. 2014). We recently demonstrated for the first time that such channels can approach the permeabilities of AQPs and CNTs while providing several advantages (Figure 7) (Licsandru et al. 2016; Shen et al. 2015). The channels tested were peptide-appended pillar[5]arene channels and imidazole-quartet artificial proton channels.

Artificial channels provide distinct advantages for scaleup when compared to CNTs and AQPs because of their compatibility with organic solvents and chemical and biological stability. They could thus be suitable for incorporation in selective high-permeability membranes.

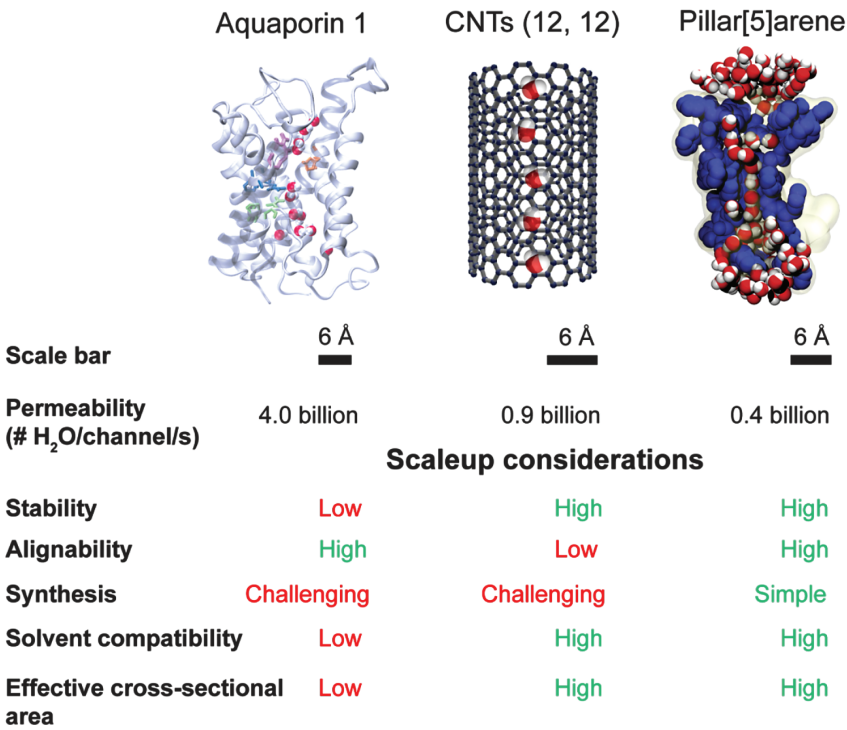

FIGURE 7 Biological water channels, aquaporins (AQPs), and their synthetic analogs, carbon nanotubes (CNTs), have high water permeabilities of ~1–10 billion water molecules per second. They have been integrated into membranes, but these membranes face scaleup challenges. We have recently shown that specific artificial water channels, peptide-appended pillar[5]arenes (PAPs), have transport rates similar to those of AQPs and CNTs. PAPs also have several advantages for scaleup, including high usable cross section, simple synthesis, organic solvent compatibility, and stability (both chemical and biological). Figure can be viewed in color at https://www.nap.edu/catalog/23659.

Graphene-based membranes can also be considered as an example of channel-based membranes and may be promising as next-generation RO membranes (Cohen-Tanugi and Grossman 2012; Mi 2014; Werber et al. 2016b). Graphene is a single thin layer of sp2 hybridized carbon that has unusual mechanical, thermal, and electrical properties and may lend itself to a variety of applications.

Pores drilled into graphene may be an option for filtration membranes but currently the pores cannot be made small enough to reject salt (Wang and Karnik 2012). More practical for desalination is the use of oxidized graphene or graphene oxide sheets stacked together so that the distance between the layers can be small enough to reject solutes (Mi 2014). This work is rapidly progressing and could be a new material for sustainable desalination.

Fouling-Resistant Membranes

A major challenge during operation of RO membranes is the deposition of colloidal materials and organic macromolecules on the membrane surface and the growth of microbes. This deposition leads to cake formation, irreversible adsorption, and growth of persistent biofilms, collectively referred to as fouling.

Fouling can cause a substantial increase in power consumption due to additional resistance to flow. In addition, salt accumulates in fouling cake layers. The cake-enhanced concentration polarization and, for biofilms, biofilm-enhanced osmotic pressure (Herzberg and Elimelech 2007; Hoek and Elimelech 2003) increase the effective osmotic pressure to be overcome, thus decreasing the driving force for membrane filtration and increasing power consumption.

Several membrane modification strategies are under consideration to reduce membrane fouling in RO systems. These include the grafting of superhydrophilic or amphiphilic molecules that can prevent adsorption of macromolecules and biological cells; use of nanoparticles, carbon-based materials such as CNTs, and graphene oxide flakes to impart biocidal properties to the RO membrane surface; and use of electroactive or magnetically actuated surfaces to prevent deposition or cause cell death. Methods that interrupt or manipulate cell-to-cell communication are also being explored for biofouling control.

Desalination Powered by Renewable Energy

Desalination has always been considered incompatible with renewable energy infrastructure because of its energy-intensive processes (Charcosset 2009). But with the rapid improvement in RO membranes and systems and concomitant decrease in energy use, more attention is being paid to the coupling of desalination units to solar (using photovoltaics) or wind energy sources. The applications are so far limited to small plants and used for “off-the-grid” applications.

CRITICAL CHALLENGES IN DESALINATION

Notwithstanding rapid progress in the development and deployment of membrane desalination in recent years, there are still persistent fundamental and practical challenges to its sustainable implementation.

Inscrutability of desalination membranes. Although crosslinked TFC RO membranes have been used for a few decades now, the microstructural details of these membranes remain unknown. This lack of knowledge prevents the establishment of a direct link between modifications in the chemistry and microstructure that drive transport properties. Efforts are ongoing to develop tools to enhance understanding of RO membrane structure.

Concentration polarization. When salt is rejected from the surface of RO membranes it forms a concentrated layer adjacent to the membrane, reducing the driving force for transport across the membrane. The thickness of this concentration polarization layer can be reduced by enhancing the back transport of solutes. Several ideas have been tested at various scales but their implementation in a sustainable manner has been challenging.

Seawater intakes and discharges. A particular challenge to the development of seawater desalination plants (including RO plants) is the impingement and entrainment of marine microorganisms during intake to the plant. Impingement is the collision and trapping of marine organisms that are larger than intake screens; entrainment is the passage of small organisms through these screens and the subsequent destruction of these marine organisms. Also, when dense brine is discharged back to the ocean, it can have detrimental effects on the marine environment if proper mixing does not occur. Efforts are needed to better understand these challenges as well as the effect of intake designs and discharge diffusers on the marine environment (Szeptycki et al. 2016).

Inland desalination brine disposal. Whereas coastal plants can discharge concentrated brine to the ocean, inland RO plants need to find a sustainable avenue to manage their brine, which could be as high as 20 percent of the feed flow. Brine minimization and beneficial reuse of brine components as sustainable alternatives to deep well disposal, disposal for municipal sewers, and use of evaporation ponds need to be evaluated carefully.

Lack of chlorine resistance in polyamide membranes. Sodium hypochlorite (i.e., bleach) is ubiquitous in water treatment plants for preventing growth of biofilms on surfaces in contact with water, including types of water treatment membranes. But this is not an option for polyamide membranes commonly used for desalination because of their high susceptibility to damage from chlorine. Development of chlorine-resistant membranes is an important practical need.

Translation of new materials. Many new materials have been developed for RO desalination, but their translation to products and use at larger scales is limited. Efforts are needed to translate innovations in materials and process design to actual products and plants.

High-salinity streams. High-salinity streams emerge from energy operations such as hydraulic fracturing (fracking), proposed underground CO_2 storage, unconventional oil development, and flue gas desulfurization applications that frequently have TDS values in excess of 100,000 ppm. These pose unique challenges to RO materials, RO process components, and operating strategies.

OUTLOOK

Membrane desalination technology is growing rapidly and becoming a critical tool for ensuring long-term water sustainability around the world. There is intense scientific interest in improving the sustainability of this technology, and several innovations are looking to further reduce the technique's power consumption and barriers to widespread use and sustainability. The future of this technology is bright, and it is expected to play a major role in the resource-limited future facing the world.

REFERENCES

Agre P. 2004. Aquaporin water channels. Angewandte Chemie 43:4278–4290.

Agre P, King LS, Yasui M, Guggino WB, Ottersen OP, Fujiyoshi Y, Engel A, Nielsen S. 2002. Aquaporin water channels: From atomic structure to clinical medicine. Journal of Physiology 542:3–16.

Al-Karaghouli A, Kazmerski LL. 2013. Energy consumption and water production cost of conventional and renewable-energy-powered desalination processes. Renewable and Sustainable Energy Reviews 24:343–356.

Australian NWC [National Water Commission]. 2008. Emerging Trends in Desalination: A Review. Waterlines Report Series No. 9. Turner, Australian Capital Territory.

Baker RW. 2004. Reverse osmosis. In: Membrane Technology and Applications, 2nd ed. Chichester, UK: John Wiley and Sons.

Barboiu M. 2012. Artificial water channels. Angewandte Chemie International Edition 51:11674–11676.

Cadotte JE. 1981. Interfacially synthesized reverse osmosis membrane. US Patent 4,277,344 A.

Charcosset C. 2009. A review of membrane processes and renewable energies for desalination. Desalination 245:214–231.

Cohen-Tanugi D, Grossman JC. 2012. Water desalination across nanoporous graphene. Nano Letters 12:3602–3608.

Cohen-Tanugi D, McGovern RK, Dave SH, Lienhard JH, Grossman JC. 2014. Quantifying the potential of ultra-permeable membranes for water desalination. Energy & Environmental Science 7:1134–1141.

de Groot BL, Grubmuller H. 2001. Water permeation across biological membranes: Mechanism and dynamics of aquaporin-1 and GlpF. Science 294:2353–2357.

Elimelech M, Phillip WA. 2011. The future of seawater desalination: Energy, technology, and the environment. Science 333:712–717.

Forbes R. 1948. A Short History of the Art of Distillation: From the Beginnings Up to the Death of Cellier Blumenthal. Leiden: EJ Brill.

Gude VG. 2011. Energy consumption and recovery in reverse osmosis. Desalination and Water Treatment 36:239–260.

Herzberg M, Elimelech M. 2007. Biofouling of reverse osmosis membranes: Role of biofilm-enhanced osmotic pressure. Journal of Membrane Science 295:11–20.

Hinds B. 2007. Molecular dynamics: A blueprint for a nanoscale pump. Nature Nanotechnology 2:673–674.

Hoek EM, Elimelech M. 2003. Cake-enhanced concentration polarization: A new fouling mechanism for salt-rejecting membranes. Environmental Science and Technology 37:5581–5588.

Israelachvili JN. 2011. Intermolecular and Surface Forces, rev 3rd ed. Burlington, MA: Academic Press.

Licsandru E, Kocsis I, Shen Y-X, Murail S, Legrand Y-M, van der Lee A, Tsai D, Baaden M, Kumar M, Barboiu M. 2016. Salt-excluding artificial water channels exhibiting enhanced dipolar water and proton translocation. Journal of the American Chemical Society 138:5403–5409.

Loeb S, Sourirajan S. 1963. Sea water demineralization by means of a semipermeable membrane. University of California Los Angeles, Department of Engineering.

Mi B. 2014. Graphene oxide membranes for ionic and molecular sieving. Science 343:740–742.

NRC [National Research Council]. 2008. Desalination: A National Perspective. Washington, DC: National Academies Press.

Reid C, Breton E. 1959. Water and ion flow across cellulosic membranes. Journal of Applied Polymer Science 1:133–143.

Shen Y-X, Saboe PO, Sines IT, Erbakan M, Kumar M. 2014. Biomimetic membranes: A review. Journal of Membrane Science 454:359–381.

Shen Y-X, Si W, Erbakan M, Decker K, De Zorzi R, Saboe PO, Kang YJ, Majd S, Butler PJ, Walz T, Kumar M. 2015. Highly permeable artificial water channels that can self-assemble into two-dimensional arrays. Proceedings of the National Academy of Sciences 112:9810–9815.

Szeptycki L, Hartge E, Ajami N, Erickson A, Heady WN, LaFeir L, Meister B, Verdone L, Koseff JR. 2016. Marine and Coastal Impacts of Ocean Desalination in California. A report of Water in the West, Center for Ocean Solutions, Monterey Bay Aquarium, and the Nature Conservancy. Available at http://waterinthewest.stanford.edu/sites/default/files/Desal_Whitepaper_FINAL.pdf.

Wang EN, Karnik R. 2012. Water desalination: Graphene cleans up water. Nature Nanotechnology 7:552–554.

Werber JR, Deshmukh A, Elimelech M. 2016a. The critical need for increased selectivity, not increased water permeability, for desalination membranes. Environmental Science and Technology Letters 3:112–120.

Werber JR, Osuji CO, Elimelech M. 2016b. Materials for next-generation desalination and water purification membranes. Nature Reviews Materials 1:16018.

Westmoreland JC. 1968. Spirally wrapped reverse osmosis membrane cell. US Patent 3,367,504 A.

Zhao Y, Qiu C, Li X, Vararattanavech A, Shen W, Torres J, Helix-Nielsen C, Wang R, Hu X, Fane AG. 2012. Synthesis of robust and high-performance aquaporin-based biomimetic membranes by interfacial polymerization-membrane preparation and RO performance characterization. Journal of Membrane Science 423:422–428.

Scalable Manufacturing of Layer-by-Layer Membranes for Water Purification

Christopher M. Stafford
National Institute of Standards and Technology

"When the well is dry, we know the worth of water."

– Benjamin Franklin

Water is critical to world health, economic development, and security. This was highlighted recently when the Obama administration hosted the White House Water Summit to raise awareness of water availability concerns across the United States and to engage stakeholders in identifying long-term solutions for water production and management suitable for investment.

BACKGROUND

Water availability is not a new issue. The demand for clean water has risen dramatically since the Industrial Revolution and will continue through the Information Age and beyond. The world's population has climbed to 7 billion, and as it expands further and water scarcity becomes a more widespread reality, it is imperative to think creatively about ways to safeguard access to clean water.

The obvious and most fundamental purpose of clean water is as a source of sustenance, to produce the food and water that every society needs to survive. Clean water is also vital to many of the complex processes that produce the technology that modern society demands and consumes. Many of those processes, however, introduce contaminants, such as heavy metals and other chemicals, into local water supplies.

For all these reasons there is a clear and growing need for technologies and processes that ensure water is clean, safe, and accessible (Shannon et al. 2008).

MEMBRANE TECHNOLOGY

Membranes and membrane technology, in particular polymer-based membranes, are key to the world's water future (Geise et al. 2010). Membranes are capable of separating a wide range of contaminants from impaired water sources, from viruses and bacteria to heavy metals to dissolved salts.

Given that water covers 71 percent of Earth's surface and 97 percent of that water is in the world's oceans, an obvious focal point of research is desalination, the recovery of water from high-salinity water sources. This can be an energy-intensive process because of the high osmotic pressure of seawater: the average sea surface salinity is 35,000 g/L (for simplicity, let's assume it is all sodium chloride), which generates an osmotic pressure ($\Delta\pi$) of nearly 400 psi or 27.4 bar. Desalination is nonetheless highly attractive because of the volume of water available for recovery.

This paper focuses on membrane desalination via reverse osmosis, so a short introduction to reverse osmosis is warranted.

REVERSE OSMOSIS

In traditional osmosis, water flows across a semipermeable membrane from regions of low solute concentrations (in this example, pure water) to regions of high solute concentration (a concentrated salt solution), in effect diluting the solute and lowering the overall free energy of the system. The driving force for the flow of water is the osmotic pressure and is dependent on the concentration of solute molecules in the concentrated solution.

In reverse osmosis, pressure is applied to the high concentration region, which has to be greater than the osmotic pressure of the solution to drive water from regions of high concentration to those of low concentration (see Figure 1), again with the aid of a semipermeable membrane. This process generates purified water on one side of the membrane and a more concentrated salt solution on the other side.

The water flux (J_w) through the semipermeable membrane can be defined as:

$$J_w = A\frac{K_w}{h}(\Delta P - \Delta\pi)$$

where A is the membrane area, K_w is the permeability of the membrane, h is the membrane thickness, and $(\Delta P - \Delta\pi)$ is the difference between the applied pressure and the osmotic pressure. From this equation, one can see that there is an inverse relationship between the applied pressure and the membrane thickness. Thus, a thinner membrane would be ideal as it would require less energy (pressure) to generate a given amount of water from an impaired water source of a given concentration of dissolved solutes (i.e., osmotic pressure).

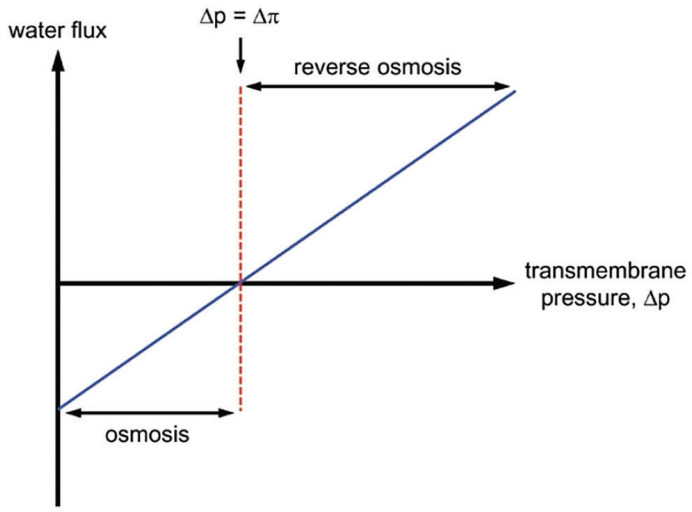

FIGURE 1 Schematic diagram of water flux as a function of applied pressure, indicating the regimes for traditional osmosis ($\Delta P < \Delta\pi$) and reverse osmosis ($\Delta P > \Delta\pi$).

PARADIGM SHIFT IN MEMBRANE TECHNOLOGY

The manufacture of today's state-of-the-art reverse osmosis membranes is based on 1970s technology of interfacial polymerization of a selective layer directly on a porous support (Cadotte 1977, 1979).

In this process, polymerization of an aromatic triacid chloride (A) and an aromatic diamine (B) occurs at the interface of two immiscible liquids, where one liquid (typically the aqueous amine solution) is wicked into the porous support. The result is a highly crosslinked, aromatic polyamide (think crosslinked Kevlar or Nomex) membrane that selectively allows the passage of water and rejects salt. The chemistry easily lends itself to roll-to-roll (R2R) or web processing, can be performed over large widths of substrates, and produces a relatively low number of defects across the membrane surface.

Over the past 40 years, this membrane technology has slowly evolved through an Edisonian, trial-and-error approach. The process makes extremely thin (100s nm) selective membranes, but they are difficult to characterize because of high roughness and large heterogeneity. Thus, understanding of how these membranes work is insufficient to allow the rational design of next-generation membranes.

In 2011 my research team at NIST proposed a paradigm shift in how these types of membranes are fabricated, in which the selective layer is created layer by layer through a reactive deposition process. We anticipated the resulting mem-

branes to be smooth, tailorable, and exceptionally thin (10s of nm). The ability to tune the membrane thickness makes this process attractive due to potential energy savings from reduced pressure requirements.

In our original demonstration (Johnson et al. 2012), we used a solution-based deposition process in which we sequentially and repeatedly layered each reactive monomer (A + B) onto a solid substrate through an automated spin coating process. We observed growth rates of approximately 0.34 nm/cycle, where one cycle represents a single (A + B) deposition sequence.

The growth rate was shown to be dependent on monomer chemistry, spin conditions, and rinse solvents (Chan et al. 2012). Additionally, the layer-by-layer films are quite smooth, exhibiting a remarkably low root mean square (RMS) roughness of 2 nm compared to commercial interfacial polymerized membranes that exhibit an RMS roughness of 100 nm or more.

The fact that the films are relatively smooth and homogeneous has two compelling advantages: (1) it enables advanced measurements of the film structure via scattering- or reflectivity-based techniques, among others, and (2) it allows *quantitative* structure-property relationships to be developed as the film thickness is well defined. X-ray photoelectron spectroscopy and swelling measurements indicate that the crosslink density of the layer-by-layer membranes is comparable to that of their commercial counterparts, even though the layer-by-layer films are considerably thinner (Chan et al. 2013).

Other researchers have adopted this approach and verified that membranes produced using this layer-by-layer process indeed have viable water flux and salt rejection (Gu et al. 2013).

TECHNOLOGICAL CHALLENGES

One major drawback to the solution-based layer-by-layer approach is throughput: spin-assisted assembly is a relatively slow process and not easily scalable.

We have started to explore the use of a vapor-based approach, in which each monomer is deposited in the gas phase, similar to atomic layer deposition of metals and oxides (Sharma et al. 2015). Each monomer/precursor is (1) heated in order to build up sufficient vapor pressure of the precursor and then (2) metered into a rotating drum reactor through dosing ports with differential pumping and purge ports on either side (see Figure 2). Again, the number of cycles (or number of consecutive ports) determines the thickness of the resulting membrane.

This approach has many advantages—such as speed, safety, and scalability—over the solution-based approach. We have shown that we can deposit 20 layers of (A + B) per minute (3 s/cycle), compared to 1 layer of (A + B) every 2 minutes using the solution-based approach (2 min/cycle). The growth rate using the vapor-based approach (0.36 nm/cycle) is nearly identical to the solution-based approach, ensuring that the processes are similar. One key advantage of the

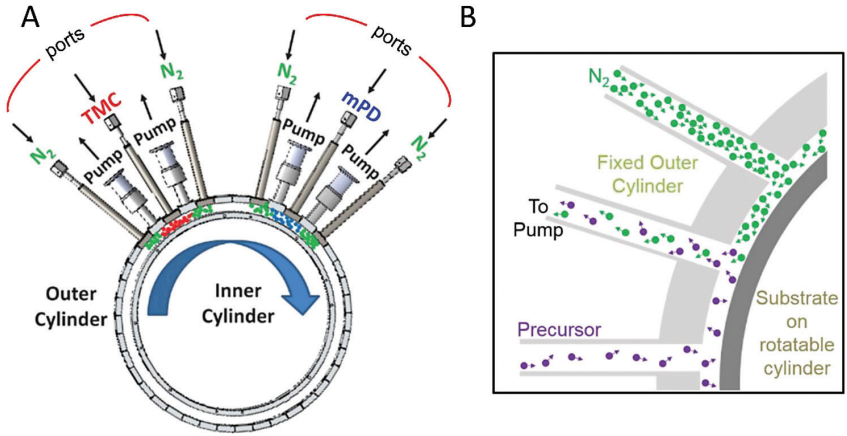

FIGURE 2 (A) Schematic of a spatial molecular layer deposition reactor for the alternating deposition of reactive monomers/precursors to form polyamide membranes. mPD = *m*-phenylenediamine; N_2 = nitrogen; TMC = trimesoyl chloride. (B) Zoomed-in view of monomer/precursor arrival to the rotating/moving substrate and removal of unreacted monomer/precursor by the sweep gas and pumping. Adapted from Sharma et al. (2015). Figure can be viewed in color at https://www.nap.edu/catalog/23659.

vapor-based approach is the potential for scale-up via continuous, roll-to-roll, or web processing.

But there are still many challenges yet to overcome, from membrane support design to membrane characterization. For example, the active layer must be coated onto a microporous support layer; thus a method for adequately preventing intrusion of the reactants into the underlying support must be devised. Also, the polyamide network topology needs to be optimized to allow the highest flux of water while maintaining adequate rejection of salt. This can be achieved through judicious monomer selection and deposition conditions.

A paradigm shift in manufacturing may lead to membranes and processes that are more energy efficient, offering one solution to the grand challenge of water security.

REFERENCES

Cadotte JE. 1977. Reverse osmosis membrane. US Patent 4,039,440.

Cadotte JE. 1979. Interfacially synthesized reverse osmosis membrane. US Patent 4,277,344.

Chan EP, Lee J-H, Chung JY, Stafford CM. 2012. An automated spin-assisted approach for molecular layer-by-layer assembly of crosslinked polymer thin films. Review of Scientific Instruments 83:114102.

Chan EP, Young AP, Lee J-H, Stafford CM. 2013. Swelling of ultrathin molecular layer-by-layer polyamide water desalination membranes. Journal of Polymer Science, Part B: Polymer Physics 51:1647–1655.

Geise GM, Lee H-S, Miller DJ, Freeman BD, McGrath JE, Paul DR. 2010. Water purification by membranes: The role of polymer science. Journal of Polymer Science, Part B: Polymer Physics 48:1685–1718.

Gu J-E, Lee S, Stafford CM, Lee JS, Choi W, Kim B-Y, Baek K-Y, Chan EP, Chung JY, Bang J, Lee J-H. 2013. Molecular layer-by-layer assembled thin-film composite membranes for water desalination. Advanced Materials 25:4778–4782.

Johnson PM, Yoon J, Kelly JY, Howarter JA, Stafford CM. 2012. Molecular layer-by-layer deposition of highly crosslinked polyamide films. Journal of Polymer Science, Part B: Polymer Physics 50:168–173.

Shannon MA, Bohn PW, Elimelech M, Georgiadis JG, Mariñas BJ, Mayes AM. 2008. Science and technology for water purification in the coming decades. Nature 452:301–310.

Sharma K, Hall RA, George SM. 2015. Spatial atomic layer deposition on flexible substrates using a modular rotating cylinder reactor. Journal of Vacuum Science and Technology A 33:01A132.

New Materials for Emerging Desalination Technologies

Baoxia Mi
University of California, Berkeley

DESALINATION AS A SOLUTION TO WATER SHORTAGE

The global water shortage caused by dwindling fresh water resources and increasing water demand, and compounded by extreme climate conditions (less precipitation), has highlighted the importance of treating unconventional waters to ensure sustainable economic and societal growth in water-stressed regions (Shannon et al. 2008).

Desalination, a process that was originally defined as the removal of salts and minerals from saline water but now includes the treatment of brackish and wastewater, likely offers a long-term strategy for augmenting water supply. This technology is widely used in many parts of the world, especially the arid Middle East. For example, Israel has been heavily relying on wastewater reuse and seawater desalination to meet much of its water needs: 86 percent of its wastewater is recycled and 60 percent of its drinking water is produced by desalination. In sharp contrast, the numbers are only 7 percent and <1 percent, respectively, for California, which regularly suffers from severe drought (Stock et al. 2015).

State-of-the-art desalination technologies include (1) thermal processes such as multistage flash and multieffect distillation and (2) membrane-based processes such as reverse osmosis (RO) and electrodialysis. The RO technology, a hydraulic pressure-driven filtration process that removes contaminants from water mainly by size exclusion and charge repulsion, accounts for around 60 percent of the market thanks to its relative advantages in capital cost, energy consumption, and ease of operation. Many other processes—forward osmosis, membrane distillation, capacitive deionization, pressure-retarded osmosis, and enhanced solar evaporation—have recently emerged as attractive alternatives in view of their

promise for reducing operational energy consumption by using sustainable energy sources such as solar, geothermal, or waste heat.

Producing water by desalination at the current development stage is more expensive than treating conventional water sources. For example, the unit cost of RO seawater desalination in the United States is now about $\$2.0/m^3$ on average (it may go down to $\$1.1/m^3$ when the technology is scaled up), compared to a typical wholesale water price of \$0.1 to $\$0.5/m^3$ (Wittholz 2008; WaterCAGov 1994). The high cost is mainly because desalination requires the removal of small, soluble contaminants (e.g., salts and inorganic/organic micropollutants such as pharmaceuticals and endocrine-disrupting compounds) that are generally not a concern in conventional water treatment. Additionally, the high salt concentration in seawater imposes a thermodynamic limit of 1.1 kWh/m^3 as the theoretical minimum energy consumption at 50 percent recovery, significantly contributing to the cost of seawater desalination.

An important goal in improving desalination technology is to separate target contaminants from water more effectively and energy-efficiently. In particular, the development of high-performance desalination membranes using emerging two-dimensional (2D) nanomaterials may revolutionize membrane-based desalination technology (Stock et al. 2015).

NEW DESALINATION MEMBRANES MADE OF 2D NANOMATERIALS

Membranes made of conventional materials (e.g., polyamide) have inherent limitations in permeability, selectivity, chemical stability, and antifouling properties, severely affecting their separation performance in desalination. Recent advances in 2D nanomaterials offer an opportunity to help overcome these limitations through the fabrication of a new class of filtration membranes for desalination.

Emerging graphene-based nanomaterials possess a unique 2D structure and highly tunable physicochemical properties as well as exceptional mechanical, electrical, and biological characteristics, all of which can be advantageously leveraged to significantly improve the separation efficiency of desalination membranes (Mi 2014). Expected to be on par with carbon nanotube membranes (Holt et al. 2006) and biomimetic aquaporin membranes (Shen et al. 2014) in terms of separation capability, graphene-based membranes are much easier to scale up thanks to both the use of graphite as a low-cost raw material and membrane synthesis via facile, scalable routes.

There are two general types of graphene-based membranes, made via very different approaches and having fundamentally different separation mechanisms. The first is a porous graphene membrane made by punching nanometer pores through the ultrathin, super-strong, and impermeable graphene monolayer, as illustrated in Figure 1(a). With its precisely controlled size and manipulated

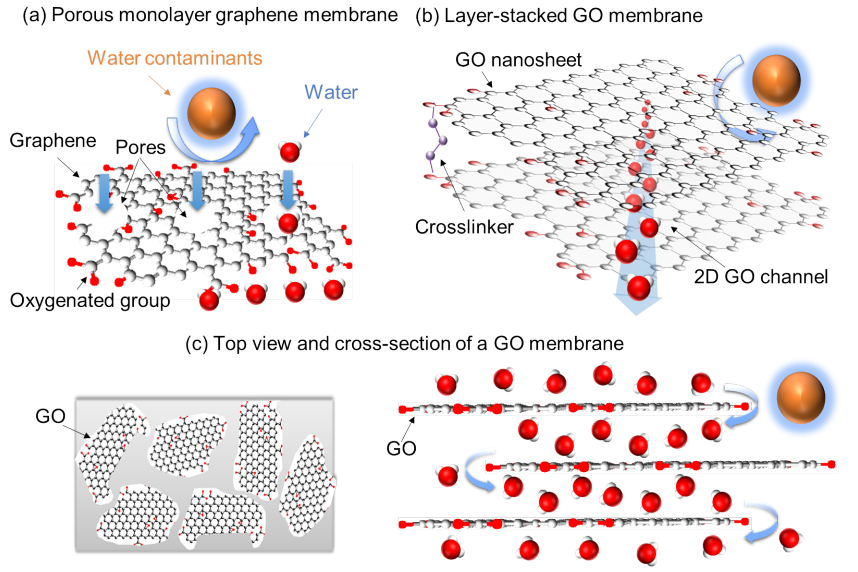

FIGURE 1 Two major types of graphene-based membranes. GO = graphene oxide. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

functional groups (which dictate the critical entrance properties) of the punched pores, the nanopore membrane allows only molecules smaller than the pores to permeate while larger molecules are rejected (Cohen-Tanugi and Grossman 2012; Suk and Aluru 2010).

In addition, the single-carbon-atom thickness (~0.3 nm) of this super-strong membrane is almost three orders of magnitude less than the thickness (typically a few hundred nanometers) of traditional desalination membranes, significantly improving the water permeability, which is inversely proportional to membrane thickness. Challenges in making such a monolayer graphene membrane include the enormous difficulties of preparing a large-area, defect-free monolayer graphene sheet and creating high-density pores of controllable, relatively uniform sizes on the graphene sheet.

The second type of graphene-based membrane is made of mass-producible graphene oxide (GO) nanosheets. As illustrated in Figure 1(b), the unique 2D structure of GO nanosheets makes it possible to synthesize a membrane via a simple, scalable layer-stacking technique (Hu and Mi 2013, 2014). The nanochannels formed between the layer-stacked GO nanosheets, functionally equivalent to

nanopores in the monolayer graphene membrane, provide a zigzag water transport path while rejecting unwanted ions and molecules that are larger than the inter-GO-layer spacing (Figure 1c). Simulation and experimental evidence indicate that, because of the very large slip length (i.e., low friction) of water molecules on a graphene surface, water can flow at an extremely high rate in the planar graphene nanochannels (Kannam et al. 2012; Nair et al. 2012), a property that could lead to the formation of highly permeable membranes for desalination.

The layer-stacking synthesis approach also enhances the adjustability of the spacing and functionalities of GO nanosheets to optimize membrane permeability and selectivity. Moreover, the 2D carbon-walled channel surface yields stronger carbon-organic interactions and thus hinders the diffusion of organic molecules in the membrane. As a result, the GO membrane can efficiently remove neutral organic contaminants (Zheng and Mi 2016), a unique feature when compared to traditional polymeric RO membranes, which are typically charged and have a relatively poor removal rate for neutral molecules.

Graphene-based nanomaterials can also be used to modify existing membranes for improved performance or multifunctionality. For example, the semiconducting property of GO nanosheets and their composites (e.g., GO–titanium dioxide) makes GO photoactive under both ultraviolet and visible lights, a useful property for developing photocatalytic membranes (Gao et al. 2014). And GO nanosheet assembly is a convenient way to form a dense barrier layer on the porous side of a traditional asymmetric membrane for fouling control in pressure-retarded osmosis, a desalination-related, energy-production process whose advancement has been hindered by the accumulation of foulants in the porous membrane support (Hu et al. 2016).

MAJOR CHALLENGES IN GO MEMBRANE DEVELOPMENT

The high water permeability of a GO membrane relies on the hypothetical existence of a continuous, nearly frictionless path for water flow in the extremely smooth graphitic (i.e., nonoxidized) regions of GO nanochannels. But heavily oxidized GO regions, which represent a large portion of the GO basal plane, do not provide a frictionless pathway and could significantly affect water flow.

As illustrated in Figure 2(a), a GO nanosheet is composed of three distinct regions: graphitic, oxidized, and defect, illustrated by a hole in the middle. The graphitic region typically occupies less than half of the total area of a GO nanosheet prepared using the Hummers method (Hummers and Offeman 1958; Marcano et al. 2010). Because graphitic regions even with the same overall area ratio could be distributed quite differently in GO nanosheets, as illustrated in Figure 2(b,c), the resulting water transport paths, boundary effects, and membrane separation capabilities can be dramatically different.

The microstructure of GO nanochannels as well as the associated water and molecular transport mechanisms are not clearly understood. Efforts are needed to

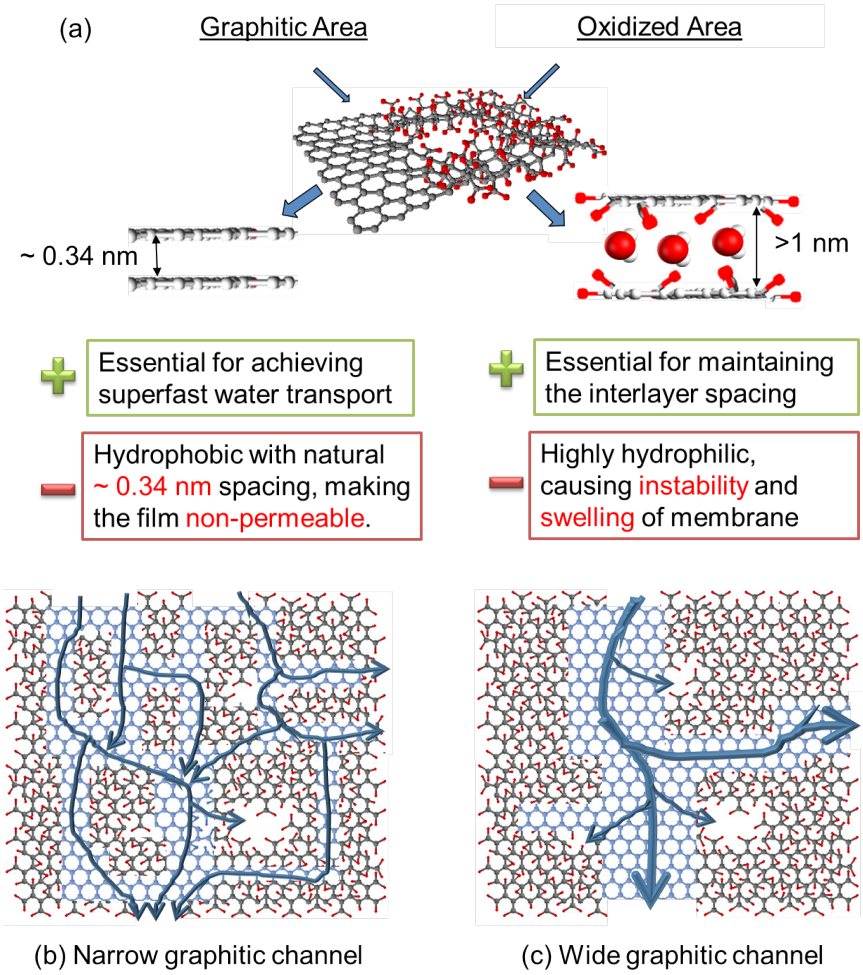

FIGURE 2 Effects of GO nanostructure on water and molecular transport. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

precisely control the size of GO nanochannels, characterize the transport length and channel width, and build mechanistic models to correlate such characteristics to membrane performance.

Controlling the interlayer spacing in a GO membrane is another critical challenge to the manufacture of effective desalination membranes. Studies have shown that it is relatively straightforward to construct a membrane with GO inter-

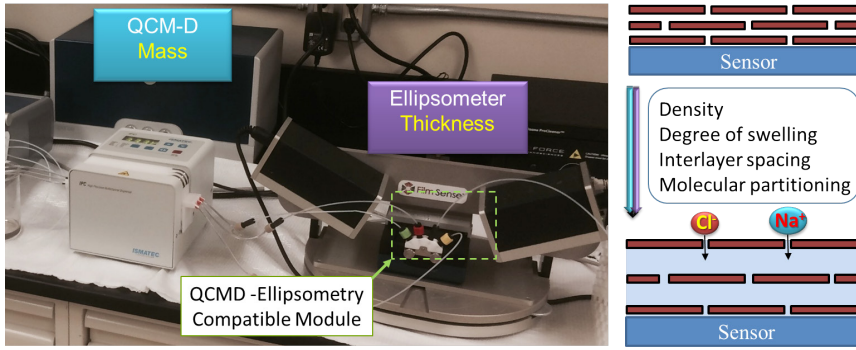

FIGURE 3 Simultaneous measurement of membrane mass and thickness by quartz crystal microbalance with dissipation (QCM-D) and ellipsometer. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

layer spacing of more than 1 nm (Hu and Mi 2013, 2014; Zheng and Mi 2016). But it becomes challenging to reduce the spacing to less than 0.8 nm (a critical value for desalination membranes to achieve high removal of sodium chloride by size exclusion) because the oxidized region in GO starts to create strong hydration forces and charge repulsion that cause membrane swelling and thus increase interlayer spacing. To accurately quantify the degree of swelling and interlayer spacing, a protocol has recently been established to simultaneously measure the mass of GO thin film by quartz crystal microbalance with dissipation (QCM-D) and film thickness by ellipsometer (Figure 3). A GO film can swell to about three times its size when it changes from dry to wet. Potential strategies to overcome such swelling include creating short covalent bonds, crosslinking out of aqueous solution, and inserting appropriately sized spacers between GO layers.

CONCLUDING REMARKS

Notwithstanding the challenges, 2D graphene-based nanomaterials hold great promise for revolutionizing membrane-based desalination technology, thanks to their advantages over traditional materials in enabling the synthesis of a desalination membrane via simple, scalable layer-stacking techniques and in the flexible manipulation of membrane permeability and selectivity for target contaminants. Other 2D materials (e.g., zeolite, molybdenum disulfide), with unique configurations that could help control interlayer spacing, are also attracting research interest in making high-performance membranes (Kang et al. 2014).

In addition, 2D nanomaterials can be innovatively constructed into a 3D

structure and thus function as a nanosized reactor to further enhance membrane selectivity and minimize membrane fouling (Jiang et al. 2015). Finally, it is worth noting that 2D nanomaterials are finding potential applications in nonmembrane-based desalination technologies. For example, 2D graphene material–enabled thin films may be used to enhance solar evaporation (Ghasemi et al. 2014) and thus help desalinate water by using sustainable solar energy.

ACKNOWLEDGMENTS

This article is based on work supported by the National Science Foundation under grant no. CBET-1565452. The opinions expressed are those of the author and do not necessarily reflect those of the sponsor.

REFERENCES

Cohen-Tanugi D, Grossman JC. 2012. Water desalination across nanoporous graphene. Nano Letters 12:3602–3608.

Gao Y, Hu M, Mi B. 2014. Membrane surface modification with tiO2-graphene oxide for enhanced photocatalytic performance. Journal of Membrane Science 455:349–356.

Ghasemi H, Ni G, Marconnet AM, Loomis J, Yerci S, Miljkovic N, Chen G. 2014. Solar steam generation by heat localization. Nature Communications 5.

Holt J, Park H, Wang Y, Stadermann M, Artyukhin A, Grigoropoulos C, Noy A, Bakajin O. 2006. Fast mass transport through sub-2-nanometer carbon nanotubes. Science 312:1034–1037.

Hu M, Mi B. 2013. Enabling graphene oxide nanosheets as water separation membranes. Environmental Science & Technology 47:3715–3723.

Hu M, Mi B. 2014. Layer-by-layer assembly of graphene oxide membranes via electrostatic interaction. Journal of Membrane Science 469:80–87.

Hu M, Zheng S, Mi B. 2016. Organic fouling of graphene oxide membranes and its implications for membrane fouling control in engineered osmosis. Environmental Science & Technology 50:685–693.

Hummers WS, Offeman RE. 1958. Preparation of graphitic oxide. Journal of the American Chemical Society 80:1339–1339.

Jiang Y, Wang W-N, Liu D, Nie Y, Li W, Wu J, Zhang F, Biswas P, Fortner JD. 2015. Engineered crumpled graphene oxide nanocomposite membrane assemblies for advanced water treatment processes. Environmental Science & Technology 49:6846–6854.

Kang Y, Emdadi L, Lee MJ, Liu D, Mi B. 2014. Layer-by-layer assembly of zeolite/polyelectrolyte nanocomposite membranes with high zeolite loading. Environmental Science & Technology Letters 1:504–509.

Kannam SK, Todd BD, Hansen JS, Daivis PJ. 2012. Slip length of water on graphene: Limitations of non-equilibrium molecular dynamics simulations. Journal of Chemical Physics 136.

Marcano DC, Kosynkin DV, Berlin JM, Sinitskii A, Sun ZZ, Slesarev A, Alemany LB, Lu W, Tour JM. 2010. Improved synthesis of graphene oxide. ACS Nano 4:4806–4814.

Mi B. 2014. Graphene oxide membranes for ionic and molecular sieving. Science 343:740–742.

Nair RR, Wu HA, Jayaram PN, Grigorieva IV, Geim AK. 2012. Unimpeded permeation of water through helium-leak-tight graphene-based membranes. Science 335:442–444.

Shannon MA, Bohn PW, Elimelech M, Georgiadis JG, Marinas BJ, Mayes AM. 2008. Science and technology for water purification in the coming decades. Nature 452:301–310.

Shen Y-X, Saboe PO, Sines IT, Erbakan M, Kumar M. 2014. Biomimetic membranes: A review. Journal of Membrane Science 454:359–381.

Stock S, Bott M, Carroll J, Escamilla F. 2015. Solutions to California's water crisis from half a world away. NBC Bay Area, November 3. Online at http://www.nbcbayareacom/investigations/Surviving-the-Drought-Solutions-to-California-Water-Crisis-From-Israel-339638362html.

Suk ME, Aluru NR. 2010. Water transport through ultrathin graphene. Journal of Physical Chemistry Letters 1:1590–1594.

WaterCAGov. Bulletin 166-4, Urban Water Use in California. August 1994. Online at http://www.water.ca.gov/historicaldocs/irwm/b166-1994/ch4.html.

Wittholz MK, O'Neill BK, Colby CB, Lewis D. 2008. Estimating the cost of desalination plants using a cost database. Desalination 229:10–20.

Zheng S, Mi B. 2016. Silica-crosslinked graphene oxide membrane and its unique capability in removing neutral organic molecules from water. Environmental Science: Water Research & Technology 2:717–725.

High-Recovery Desalination and Water Treatment

KEVIN L. ALEXANDER
Hazen and Sawyer

There are many water supply challenges facing government, municipal, commercial, and industrial industries throughout the United States and around the world. As demand for water continues to increase and conventional water supplies are depleted, alternative water supplies are being developed. The processes to treat the recovered waters must create a water supply that lasts well into the future without a legacy of environmental challenges.

Alternative water supplies being considered include low quality surface water, irrigation runoff, brackish groundwater, municipal and industrial wastewater, and seawater. These water sources can be low quality and substantially more expensive to treat than conventional water sources, but when the cost of treating the alternative source drops below that of the currently available source, opportunity is created.

The production of high-quality water often requires implementation of desalination technologies, which use energy in the form of pressure, heat, or electricity to remove salt from water, resulting in high energy consumption and production of a high-salinity waste stream. Solutions to these challenges may be found in new technologies and combinations of new and conventional technologies.

BACKGROUND

Thermal (i.e., distillation) processes have been used for desalination since the 1960s and remain prevalent in areas that have cheap energy sources or waste heat. Although there have been advances in thermal technologies, including multistage distillation and multistage flash distillation, thermal processes are still more expensive than reverse osmosis (RO) membrane technology.

RO desalination technology, which was incubated at universities with funding from the US government, is used in many industries that require high-quality and low-salinity water. The first commercial RO desalination project in 1971 was to produce less than a few thousand gallons per day for Texas Instruments. Today, projects requiring well over 100 million gallons per day (MGD) (378.5 megaliters per day, MLD) are being implemented and others over 1 billion gallons per day (3.785 billion liters per day) are being considered.

Advances in the field of desalination include improvements in fundamental materials, manufacturing techniques, packaging techniques, and mechanical energy recovery techniques. In most cases, the objective has been to improve operational aspects of the desalination technology to achieve maximum water recoveries at the lowest energies possible while minimizing high-salinity waste flows. Improvements have reduced temperatures, operating pressures, and electrical parameters to near theoretical levels for applications in many types of water supplies. Other advances include treatment chemicals that allow for additional water recovery by extending the saturation limits of salts in solution.

Ideally, the salt and dissolved constituents remaining after the water is separated from the source water are highly concentrated or in solid form. Achieving high water recovery minimizes the high-salinity waste stream that must be discharged. With an increasingly stringent regulatory environment, many areas of the United States will not allow discharge of these streams even to wastewater treatment facilities.

The technologies being developed to take advantage of the opportunity costs that are available with the rising cost of water (due to limited supplies) vary in their approach to minimize both brine streams and energy requirements. They include controlled-scaling RO, closed-circuit desalination, forward osmosis, electrodialysis metathesis, membrane distillation, capacitive deionization, and brine-bulb technologies. Some of these technologies result from incremental improvements to existing technologies while others represent entirely new concepts.

This paper reviews these technologies and applications. It defines the opportunities and opportunity costs to show the major drivers in the market that are attracting investors and inventors alike to the field of desalination technologies.

HOOVER DAM: A WATER SUPPLY SOLUTION EXAMPLE FOR THE NEXT GENERATION

In the 1920s and 1930s, during the Great Depression, the US government saw an opportunity to get the economy moving again by building the Hoover Dam and other major infrastructure projects along the Colorado River. The government foresaw the need for more water supply in the major urban and agricultural areas developing from Colorado through California and determined that taming the Colorado River would provide such a supply. Construction of the Hoover

Dam enabled the region to grow exponentially and contributed significantly to its economy and security.

The Hoover Dam project provides a stable and sustainable source of fresh water within and beyond the contributing watershed. Although it has brought cross-state and cross-border supply challenges, the Hoover Dam and others on the Colorado River are great examples of managed water supply solutions. The project also ensures lower total dissolved solids (TDS) water for the region, a further benefit compared to brackish sources used throughout the lower Colorado River states. Although the system of dams has been criticized for its environmental impact on the Colorado River, the advantages from power supplied, water storage, and social/recreational and economic benefits have been significant. However, in recent years the project's storage and drought mitigation capacity are being tested as demand for water increases and it becomes clear in the drought-stricken region that there is not as much water available as was originally predicted.

Today's challenges require leaders, planners, and researchers to look to the next 100 years and assess the opportunities and risks of water supply solutions. Desalination technologies, when looked at in this context, may be the Hoover Dam of this and future generations. They provide access to potential water supplies such as brackish groundwater, wastewater, and seawater, sources not typically considered because cheaper and more abundant sources were available. And they allow for immediate access to water in aquifers and the ocean and to wastewater, all of which are drought-proof sources. Water supply solutions of this generation must be able to defend against drought and unpredictable weather. One of the most important considerations when looking at long-term solutions that use lower-quality water sources is the ability to remove constituents and contaminants to levels that make the water quality commensurate with the proposed uses.

Desalination technology has challenges that affect the environment, such as higher energy consumption, which translates into greenhouse gas emissions. The technology also creates a residual brine or concentrate waste stream that can have an adverse impact on the environment if not managed properly.

A BRIEF HISTORY OF DESALINATION

The desalination technology of today was conceived in the 1960s, when President Lyndon Johnson supported Israel in the development of desalination through the US Department of the Interior's Office of Saline Water. The Israelis invested effort in technology using freezing in a vacuum, but the technology was not commercialized and did not receive widespread acceptance. However, from the research and development efforts two technologies were developed: multistage flash distillation (MSF) and multieffect distillation (MED) technology. They are still in use in desalination projects around the world.

While Israel was developing desalination technology, the US government was supporting the development of reverse osmosis. Sidney Loeb at UCLA had

developed a semipermeable cellulose acetate RO membrane that was capable of removing salt from water. The challenge was in commercializing the technology. The US government saw an opportunity and hired General Atomic to develop a commercial product using RO membranes. In 1966 ROGA, a division of General Atomic, hired Richard G. Sudak to develop the technology for commercial applications. ROGA developed the first commercial applications and in 1971 sold the first RO system to Texas Instruments. Since then, both thermal and RO desalination technologies have seen widespread application and improvements.

DESALINATION TECHNOLOGIES TODAY

Desalination technology has been used in many fields to achieve specific water quality. For example, in the power industry distillation technology is used to generate service water and potable water. In the oil refining industry, RO membranes are used to generate 2,000-pound boiler feed water, which requires very low hardness and silica. In the field of microchip manufacturing, RO membranes are used to generate 18 megohm water for specialized washing. And for municipal water, RO membranes are used for treating wastewater to meet drinking water standards for applications in indirect and (soon) direct potable water supply.

The most significant advance in RO technology has been improvements in the membrane material. In the early 1990s the industry moved away from cellulose acetate to a polyamide composite, which reduced energy requirements from 300–400 psi for brackish water applications to near theoretical osmotic pressures of 100–200 psi. The main disadvantage of the latest membrane material is that it is not oxidant tolerant and is therefore more susceptible to fouling during operation.

RO membrane manufacturing changed to automated processes starting in early 2000. Prior to that, membranes were manufactured by hand gluing and rolling. Hand rolling was challenged by quality control and a loss in membrane area. Automated rolling and manufacturing have increased the membrane area by greater than 10 percent within the systems. Automated rolling also allowed for packaging changes from 8" to 16" and 18" diameter membrane elements.

The major advances for the MSF and MED technologies have been in materials improvements to address corrosion and heat transfer. A wide range of materials can now resist corrosion and are more economical, reducing the energy consumption of the technology. In addition, because the technologies work with vacuum, they have addressed operational challenges with the aid of better sealing technology.

Energy recovery devices and approaches to energy optimization in desalination have provided significant opportunity. In recent years, there have been more energy recovery devices introduced into the market, including high-efficiency pressure exchange devices that use corrosion-resistant ceramic materials. The energy consumption for seawater desalination using RO with energy recovery has declined from 11–14 kWh per 1,000 gallons (kgal) to 8–11 kWh/kgal. In

the MSF and MED technology, the energy consumption associated with process improvements and materials has decreased from 15 kWh/kgal to 13 kWh/kgal.

DESALINATION CHALLENGES

Challenges remain in efforts to achieve the objective of desalination technology: to maximize water recovery while minimizing high-salinity waste stream and at the lowest possible energy consumption.

Challenge 1. The amount and type of salt in the water are directly proportional to how much water can be recovered. As a salt solution becomes concentrated by the removal of water, it approaches saturation, with the liquid salt crystalizing and becoming a solid. The crystal formation can form a scale on the equipment and damage the RO membranes or the equipment. The difficulty in predicting when crystal formation and scaling will occur is compounded when there are significant numbers of different types of salts in a solution and each has a different saturation level. Controlling the scaling and crystal formation becomes a major treatment challenge, with hours spent analyzing the water quality and treatment technology to determine whether and where scaling could occur.

Challenge 2. The higher the salinity, the more pressure or electrical energy required to remove the salt from the solution. For seawater at around 35,000 mg/L of salinity, the theoretical pressure required to overcome the osmotic pressure and reverse the flow of water through the membrane from the higher saline side to the clean or fresh water side of the membrane is approximately 700 psi. For sewage in a normal wastewater plant, the pressure required to overcome osmotic pressure is approximately 20–30 psi.

Challenge 3. The remaining high-salinity waste flow must be handled as part of the overall treatment process, but there are not many practical places to discharge the waste stream. Historically, the most likely discharge locations have been back to the ocean for seawater desalination plants on the coast, into streams and lakes where there are higher volumes of fresh water to dilute the flow in inland areas, into sewers for eventual treatment in sewage treatment plants, and in some locations such as Florida there is the possibility of injecting the waste flow back into the ground through injection wells. In the right environments, such as the arid regions of the country, there are opportunities for evaporation and enhanced evaporation. Unfortunately, with most of the options, the remaining water in the high-salinity waste stream cannot be recovered as a potential water source.

Limited locations and the inability to discharge flows have made desalination a challenge. However, with the ever increasing cost of water where supplies are limited, there are opportunities. For example, current and projected cost of importing and treating surface water in San Diego are projected to be $1,926 per acre foot ($6.13/kgal) by 2021. The cost of desalinating seawater is currently between

$1,500 and $3,000 per acre foot ($4.60 and $9.20/kgal) depending on location and project specific considerations along the California coast.

The cost of seawater is increasing as well due to environmental, permitting, and other concerns. However, treating sewage to drinking water is much lower, at less than $1,000 per acre foot ($3.06/kgal), and treating groundwater in the Riverside area is around $625 per acre foot ($1.90/kgal).

NEW DESALINATION TECHNOLOGIES

Because the cost of water is rising rapidly, there is opportunity for desalination technology development to allow for the production of water at a lower cost than the projected cost of importing or treating seawater in areas such as southern California. In the desalination industry, companies such as GE and venture capitalists are exploring the market and investing to capitalize on the opportunity.

Investment in research on new and improved technologies for recovering water from impaired water sources has led to many different approaches to treating the water. Some technologies that show promise and are gaining acceptance and experience are described below:

Controlled scaling RO (CSRO) allows a third or fourth stage of RO to treat the concentrate from a primary RO system. The technology operates in the final stage of the RO system beyond the theoretical saturation and uses cleaning chemicals to remove scale from the membranes to restore performance. The system is operated beyond saturation for some constituents such as calcium carbonate, calcium sulfate, silica, and potentially others. CSRO is operated in a forward and reversing operation to control scaling and membrane life. The system is cleaned on a frequent basis with various cleaners, acids, and bases to keep the membranes operating. This type of system is used at Water Replenishment District of Southern California.

Desalitech™ uses closed circuit desalination or batch desalination. The high-salinity waste stream is recirculated through the RO unit, allowing for recoveries on a batch basis as high as 97 percent. This technology uses conventional RO equipment operated in a different configuration. At the end of a batch the high-salinity waste stream is discharged and fresh water filled into the feed tank and recirculated. The system operates on low-TDS water conditions that are typical for inland desalination and wastewater desalination projects. The technology has some limitations in comparison to straight RO, such as varying water quality from the beginning to the end of each batch.

Forward osmosis uses a high-salinity stream as a draw solution to pull water from lower-salinity feed water. The feed water can be from a number of sources such as wastewater effluent, high-salinity waste streams from desalination systems, brackish well water, or other high-salinity sources as long as the salinity of the feed is much lower than the draw solution. The difference in salinity between

the feed and draw solutions provides the energy to move the water across the membrane. The draw solution in some cases is a special solution that can be separated from the water. Alternatively, this technology could be used in a seawater application, with ocean water drawing fresh water from a brackish water source, followed by dilution of the seawater to reduce salinity, reducing the feed pressure and energy required to desalinate the ocean water. This could be a viable way to treat seawater using wastewater as a source of pure water while improving the economics and environmental impact.

Brine bulb technology uses AC current across a brine stream to generate heat for evaporation under a vacuum condition, using various technologies to improve the efficiency of the water recovery in a batch process. The technology combines electrocoagulation and vapor removal to separate the salt from the solution. The benefit of the system is that it allows for further recovery of the water.

Dewvaporation, developed at Arizona State University, is a specific process of humidification-dehumidification desalination (patented as AltelaRain®) that uses air as a carrier-gas to evaporate water from saline feeds and form pure condensate at constant atmospheric pressure. The heat needed for evaporation is supplied by the heat released by dew condensation on opposite sides of a heat transfer wall. Because external heat is needed to establish a temperature difference across the wall, and because the temperature of the external heat is variable, the external heat source can be from waste heat, such as solar collectors or fuel combustion. The unit is constructed of thin wettable plastics and operated at atmospheric pressure. The technology is currently sold in the oil field but could have applications in concentrate treatment.

Other desalination technologies on the cutting edge include capacitive deionization and membrane distillation. Companies such as GE are investing in their Aquasel desalination technology, high-efficiency RO(HERO), and electrodialysis metathesis. All of these promising technologies are being tested in various applications.

SOLUTIONS FOR FUTURE GENERATIONS

The Hoover Dam solved many water supply issues for future generations. It provided a water source and energy supply source of immense volume that enabled unfettered growth in the western United States. At the time it also improved the regional economy with all of the jobs associated with the project, and that economic boost continues with jobs, tourism, energy, and water storage. The dam and the reservoir it created, Lake Mead, help to ensure a water quality balance for the region. The effects of saltwater tributaries to the Colorado River are mitigated by stored water, reducing the salinity for lower Colorado River users.

When the dam and lake were created, they offered a drought-proof water

supply. Today the limits of the water supply have been reached and the Lake is being impacted by climate change and extended drought conditions in the region.

In today's ever changing water landscape, leaders in the water industry need to consider the next generations and what solutions today will solve water supply and water quality challenges long into the future. Desalination coupled with high-recovery technologies are the tools that will be used to solve these challenges. They provide access to the next available water supplies and to water supplies once considered impossible to utilize, ensuring safe and reliable water supplies. They enable access to drought-proof water supplies such as wastewater effluent and irrigation drainage flows, and secure the region against the effects of climate change and extended drought. They require educated, trained, and skilled labor and thus yield economic and social benefits for communities. Last, the technology can offer a stable solution within the cost boundaries of the ever increasing cost of local supplies.

Desalination technologies, not unlike the Hoover Dam, are capable of solving water supply challenges. High-recovery applications are being investigated and will likely provide a path toward successful implementation of large-scale desalination in arid and inland regions with limited ability to discharge concentrate streams.

This is an exciting time in the industry to look forward and at the same time consider the past risks and rewards of visionary thinking.

Technologies for Understanding and Treating Cancer

Technologies for Understanding and Treating Cancer

Julie Champion
Georgia Institute of Technology

Peter Tessier
Rensselaer Polytechnic Institute

Cancer is a complex group of more than 100 diseases characterized by uncontrolled cell growth. Approximately 40 percent of people will be diagnosed with a form of cancer in their lifetime, and about 15 percent of all people will die from cancer. This has motivated a significant amount of research to better understand and treat these devastating diseases.

At the molecular level, cancer is caused by mutations in genes that regulate several important cellular functions. These genetic mutations may be inherited, acquired via exposure to environmental hazards (e.g., chemicals in tobacco smoke, radiation, sunlight), or caused by unknown factors.

Human cells normally grow and divide as needed and die when they are old, damaged, or overcrowded. Cancerous cells fail to respond to normal signals that regulate cell growth and death, leading to uncontrolled growth. In some cases such growth leads to the formation of large cellular masses (tumors), in others it does not (as in cancers of the blood). Malignant tumors can spread into nearby tissues, and cancerous cells can break off tumors and travel to other parts of the body to initiate the formation of new tumors.

The ability of cancer cells to coopt normal cells to form blood vessels to feed tumors and remove their waste is critical to sustaining their uncontrolled growth. Another key is their ability to evade the immune system that normally eliminates damaged or abnormal cells from the body.

Cancer presents a number of challenges that engineers from different disciplines are working to address. Understanding how cancer develops and what makes some cancer cells migrate to new sites is essential to identify the necessary conditions for these events and how they may be prevented or arrested. Early detection of cancer is known to be an important factor in survival, but more sensi-

tive and selective tools are needed to identify rare cancer cells and biomolecules indicative of cancer from highly complex biological mixtures such as blood.

Treatment of cancer also has many challenges, including high toxicity in healthy tissues, development of drug resistance, and the need to better match drugs with particular cancer subtypes. New methods of drug delivery to specifically target cancer cells and alternative therapeutic approaches with new molecules and/or physical ablation methods are needed. Additionally, better imaging methods are necessary to identify smaller tumors, assist surgeons in completely and specifically removing cancerous cells, and track response to treatment. These challenges require biological and molecular expertise together with engineering innovation.

The first speaker, Cynthia Reinhart-King (Cornell University), set the stage by discussing how cancer cells go awry. She explained how extracellular signals and the microenvironment around cancer cells influence their uncontrolled growth and expansion.

Brian Kirby (Cornell University) then addressed cancer detection. He reviewed recent advances in detecting rare cancer cells using microfluidics that can be used for noninvasive detection and improved diagnosis and treatment planning.[1]

Next, Jennifer Cochran (Stanford University) described methods for interfering with the spread of cancer—specifically, therapeutic molecules that block the ability of cancer cells to leave the initial tumor and start new ones.

Finally, Darrell Irvine (MIT) discussed strategies for harnessing the immune system to target and destroy cancer cells. He highlighted approaches that use materials science and biotechnology methods to control and sustain antitumor immune responses specific for different types of cancer.

[1] Paper not included in this volume.

How Cancer Cells Go Awry: The Role of Mechanobiology in Cancer Research

Cynthia A. Reinhart-King
Cornell University

Cancer is the second leading cause of death in the United States and is projected to overtake cardiovascular disease as the leading cause of death in the next few years. Few patients die from primary tumors, but once a tumor has spread to other parts of the body (a process called metastasis), it becomes much more difficult to treat—90 percent of cancer deaths are due to metastasis.

The exact mechanisms by which tumors form, grow, and spread are not clear, but significant attention has been paid to the role of genetic mutations in cells that drive uncontrolled growth. While genetics and gene mutations are clearly drivers of cancer, it is now known that they are not the only key players. The chemical and physical environment surrounding tumor cells also contributes to malignancy and metastasis.

Numerous tumors are diagnosed based on their physical properties; as an example, changes in tissue stiffness and density are markers of tumor formation detectable by palpation and medical imaging. Notably, changes in tissue stiffness and density have been shown to enhance tumor progression. As such, research now focuses not only on the causes and treatments of genetic mutations and molecular changes in the cell but also on physical changes in the tissue and cells. This requires a new arsenal of tools aimed at characterizing and controlling the physical properties of cells and tissue. Engineers have made significant strides in developing the tools and models necessary to understand and attack cancer.

OVERVIEW: CANCER, METASTASIS, AND STAGING

Tumors are generally thought to form from one initial rogue cell that undergoes genetic changes that result in its uncontrolled growth and proliferation (Hahn

and Weinberg 2002). As this proliferation occurs, the tumor is considered benign as long as the cell mass remains in the tissue in which it formed. In this case, it is not considered cancer, but it can sometimes be dangerous if its size compresses nerves, arteries, or other tissues. If, however, the cells invade the surrounding tissue, it is considered cancerous. The cells can spread to surrounding tissues, often traveling through the bloodstream and/or lymph systems to other organs in the body. As such, there is immense interest in determining

(1) What triggers the initiating steps that lead to uncontrolled proliferation?
(2) What are the determinants of invasion? and
(3) How can the spread of tumor cells in the lymph system and blood stream be prevented?

Staging of cancer is done to categorize the extent of the spread of the tumor in a common clinical language that all physicians can understand and use to establish a prognosis, determine a course of action, and determine the fit for a clinical trial. It is based on factors such as the location and size of the primary tumor, and whether the tumor has spread to the lymph nodes or other areas of the body. The TNM staging system is based on categorizing the size and extent of the primary tumor (T), the spread to regional lymph nodes (N), and the presence of secondary metastatic tumors in other organs (M). Each of these (T, N, and M) is then used to determine the numerical stage of the cancer (I–IV).

Staging is specific to cancer type and depends on the type of tumor and its location. Stage I is indicative of a cancer that is the least advanced and has the best prognosis; Stage IV indicates that the cancer has spread to other areas of the body and is typically much more difficult to treat. In general, if a cancer is identified and treated before it has spread, the outcomes are favorable. For this reason, significant attention has been paid to stopping cancer progression before the cells metastasize.

MECHANOBIOLOGY IN CANCER

Most cancer research has focused on identifying genetic drivers and understanding the mechanism by which genes and specific signaling pathways in the cell drive tumor formation. Notably, however, cancer, from tumor initiation through metastasis and the formation of secondary tumors, involves both genetic changes in a cell and physical changes to both the tissue structure and the cancer cells (Carey et al. 2012). More recent work has therefore also focused on the mechanobiology of tumor progression. Mechanobiology is the study of how forces (e.g., pressure, tension, and fluid flow) and mechanical properties (e.g., stiffness and elasticity) affect cellular function.

Recent advances in engineering and the physical sciences have uncovered critical roles of the mechanical and structural properties of cells and tissues in

guiding malignancy and metastasis. Indeed, it is increasingly appreciated that tissue architecture and the mechanical properties of tissues and cells contribute to cancer progression.

The ability of cells to exert force against their surroundings, as one example, enables the rearrangement of tissue fibers and the creation of paths through the tissue that facilitate metastatic cell movements (Kraning-Rush et al. 2011, 2013). The ability to generate these forces is enhanced in tumor cells compared to their healthy counterparts (Kraning-Rush et al. 2012), suggesting that metastatic cells are better at invading tissue because, in part, they are better at physically rearranging it to create paths in which they can move.

Metastatic cells have also been shown to be more deformable than nonmetastatic cells (Agus et al. 2013; Guck et al. 2005). This deformability may help metastatic cells squeeze through tissue to escape the primary tumor and enter the circulatory system to move to secondary sites.

In addition to changes in the physical properties of the cells, changes in the physical properties of the tissue have been shown to promote cancer progression. Solid tumor tissue is stiffer than normal tissue, and research suggests that this stiffness can promote tumor cell growth and invasion. Conversely, decreasing tissue stiffness has been shown to delay tumor progression (Cox et al. 2016; Venning et al. 2015). These results indicate that tissue mechanics plays a critical role in cancer and that, clinically, approaches to intervene with cancer mechanobiology may benefit cancer treatment.

TISSUE-ENGINEERED PLATFORMS TO STUDY MECHANOBIOLOGY IN CANCER

To investigate and manipulate the mechanobiology of cancer, tissue-engineered platforms have been critical. Because it is known that there are distinct differences between how tumors form and grow in the human body compared to how cancer cells grow in culture dishes or in animal models, tissue-engineered platforms serve as bridges to better understand and manipulate the drivers of cancer.

Using principles from biomaterials science, mechanics, and chemistry, engineers have been working to create platforms that recreate the architecture, chemistry, and mechanical properties of the tumor microenvironment (Carey et al. 2012; Mason and Reinhart-King 2013). These platforms will enhance understanding of the physical forces and features that drive tumor progression, and have also in many cases been adapted for use in drug testing.

Tissue-engineered platforms can be created to mimic both the dimensionality of tumors, overcoming the limitation of conventional cell culture, and the stiffness and porosity of native tissue at various stages of tumor progression. More specifically, several bioengineering groups have developed tunable materials that mimic the changing mechanical properties of the tumor microenvironment. Materials that

can be activated to stiffen or soften through various chemical techniques have been created and used to study how cells respond to the dynamic mechanical environment in tumor tissue (Gill et al. 2012; Magin et al. 2016). They may be useful for parsing the effects of genetic changes from those induced by changes in the mechanical environment of the cell.

TRANSLATING MECHANOBIOLOGY TO THE CLINIC

One of the biggest hurdles in the field of cancer mechanobiology is the translation of findings to the clinic. For instance, because it is known that metastatic cells exert higher forces and are more deformable than nonmetastatic cells, there may be clinical value to assaying patient samples to correlate forces and deformability with patient prognosis. It has been suggested that new mechanobiological assays be incorporated in clinical protocols, and significant efforts are being made to develop assays that are user-friendly and translatable to clinical settings (Kiessling et al. 2013).

Clinically targeting mechanically related molecules may also be feasible. There are numerous signaling pathways and associated proteins in cells that control cellular force profiles, cell contractility, and cell deformability. In fact, many of these pathways have been either directly or indirectly pharmacologically targeted for the treatment of other diseases, including arthritis, diabetes, cardiovascular disease, and pulmonary diseases.

Additionally, approaches to alter the mechanical properties of tissues have been developed for targeting tissue stiffening in wound healing and cardiovascular disease. Thus clinically treating changes in the mechanical properties of cell and tissues is feasible and within reach.

SUMMARY

The field of cancer mechanobiology has grown significantly over the past decade as the role of mechanical forces in cancer has been increasingly appreciated. It is now well accepted that mechanical changes in both cells and tissues can contribute to malignancy and metastasis, but the mechanisms by which these changes promote cancer are not yet fully understood. Engineers have the unique skills to build platforms to measure, probe, and manipulate cell and tissue mechanics to better understand cancer mechanobiology and translate it to the clinic.

REFERENCES

Agus DB, Alexander JF, Arap W, Ashili S, Aslan JE, Austin RH, Backman V, Bethel KJ, Bonneau R, Chen WC, and 85 others. 2013. A physical sciences network characterization of non-tumorigenic and metastatic cells. Scientific Reports 3:1449.

Carey SP, D'Alfonso TM, Shin SJ, Reinhart-King CA. 2012. Mechanobiology of tumor invasion: Engineering meets oncology. Critical Reviews in Oncology/Hematology 83:170–183.

Cox TR, Gartland A, Erler JT. 2016. Lysyl oxidase, a targetable secreted molecule involved in cancer metastasis. Cancer Research 76:188–192.

Gill BJ, Gibbons DL, Roudsari LC, Saik JE, Rizvi ZH, Roybal JD, Kurie JM, West JL. 2012. A synthetic matrix with independently tunable biochemistry and mechanical properties to study epithelial morphogenesis and EMT in a lung adenocarcinoma model. Cancer Research 72:6013–6023.

Guck J, Schinkinger S, Lincoln B, Wottawah F, Ebert S, Romeyke M, Lenz D, Erickson HM, Ananthakrishnan R, Mitchell D, and 3 others. 2005. Optical deformability as an inherent cell marker for testing malignant transformation and metastatic competence. Biophysical Journal 88:3689–3698.

Hahn WC, Weinberg RA. 2002. Mechanisms of disease: Rules for making human tumor cells. New England Journal of Medicine 347:1593–1603.

Kiessling TR, Herrera M, Nnetu KD, Balzer EM, Girvan M, Fritsch AW, Martin SS, Kas JA, Losert W. 2013. Analysis of multiple physical parameters for mechanical phenotyping of living cells. European Biophysics Journal 42:383–394.

Kraning-Rush CM, Carey SP, Califano JP, Smith BN, Reinhart-King CA. 2011. The role of the cytoskeleton in cellular force generation in 2D and 3D environments. Physical Biology 8:015009.

Kraning-Rush CM, Califano JP, Reinhart-King CA. 2012. Cellular traction stresses increase with increasing metastatic potential. PLoS One 7:e32572.

Kraning-Rush CM, Carey SP, Lampi MC, Reinhart-King CA. 2013. Microfabricated collagen tracks facilitate single cell metastatic invasion in 3D. Integrative Biology 5:606–616.

Magin CM, Alge DL, Anseth KS. 2016. Bio-inspired 3D microenvironments: A new dimension in tissue engineering. Biomedical Materials 11:022001.

Mason BN, Reinhart-King CA. 2013. Controlling the mechanical properties of three-dimensional matrices via non-enzymatic collagen glycation. Organogenesis 9:70–75.

Venning FA, Wullkopf L, Erler JT. 2015. Targeting ECM disrupts cancer progression. Frontiers in Oncology 5:224.

Engineered Proteins for Visualizing and Treating Cancer

Jennifer R. Cochran
Stanford University

Cancer is complex and its diagnosis and treatment can more effectively be tackled by teams of scientists, engineers, and clinicians whose expertise spans bench-to-bedside approaches.

An emerging core philosophy applies understanding of molecular mechanisms underlying disease pathophysiology as design criteria toward the development of safer and more efficacious tumor targeting agents (Kariolis et al. 2013). Armed with this knowledge, academic and industrial researchers are using a variety of approaches to create tailor-made proteins for application in cancer imaging and therapy. These efforts leverage enabling tools and technologies, including methods for (1) protein design and engineering, (2) biochemical and biophysical analyses, and (3) preclinical evaluation in animal models.

Important deliverables of this work include insight into ligand-mediated cell surface receptor interactions that drive disease, and the development of new protein-based drugs and imaging agents for translation to the clinic.

BACKGROUND

As the field of protein engineering evolved during the 1980s, modified proteins soon joined recombinant versions of natural proteins as a major class of new therapeutics. The ability to customize the biochemical and biophysical properties of proteins to augment their clinical potential has presented many exciting new opportunities for the pharmaceutical industry.

The market value of such biopharmaceuticals is currently more than $140 billion, exceeding the GDP of three-quarters of the economies in the World Bank rankings (Walsh 2014). Monoclonal antibodies used to treat cancer, rheumatoid

arthritis, and cardiovascular and other diseases account for a large share of these efforts (Drewe and Powell 2002). In 2014 the US and European markets included close to 50 monoclonal antibody drugs, a $75 billion market (Ecker et al. 2015). In 2015, the top three revenue-generating cancer drugs were monoclonal antibodies: rituximab (Rituxan®), bevacizumab (Avastin®), and trastuzumab (Herceptin®), all produced by Genentech/Roche. The size of this market underscores both the clinical and economic importance of protein therapeutics in translational medicine.

CURRENT CHALLENGES

Challenges for cancer therapeutics include the need for more selective localization to tumors versus healthy tissue, and improved tissue penetration and delivery to brain tumors, which are protected by the restrictive blood-brain barrier. Other therapeutic challenges are tumor heterogeneity that makes cancers difficult to treat, acquired drug resistance that cannot be overcome because of dose limiting drug toxicity, and lack of effective drugs to treat cancer once it has spread.

Limitations of monoclonal antibodies in addressing these and other challenges have motivated the development of alternative tumor targeting proteins with different molecular sizes and biophysical attributes, conferring altered pharmacological properties (Weidle et al. 2013). In the following sections I describe some examples of engineered protein therapeutics developed by our research team that have opportunities to affect cancer in new and impactful ways.

AN ULTRA-HIGH AFFINITY ENGINEERED PROTEIN THERAPEUTIC FOR TREATING METASTATIC DISEASE

Despite advances over the past few decades in the development of targeted therapeutics, there is a lack of effective drugs to treat cancers once they have spread (called metastasis), and 90 percent of patients succumb to metastatic disease. We teamed up with cancer biologist Amato Giaccia (Stanford Radiation Oncology) to address this challenge.

In a number of human cancers, aberrant signaling through the Axl receptor tyrosine kinase has been demonstrated to drive metastasis (Li et al. 2009), confer therapeutic resistance (Hong et al. 2013), and promote disease progression (Vajkoczy et al. 2006). Additionally, Axl overexpression has been observed in multiple solid and hematological malignancies (Linger et al. 2008), with expression levels often correlating with disease stage and poor clinical prognosis (Gustafsson et al. 2009; Hong et al. 2013; Rankin et al. 2010). Ambiguity surrounding the fundamental characteristics of Axl's interaction with its ligand, growth arrest-specific 6 (Gas6), including its affinity and the mechanism of receptor activation, have hindered the development of effective Axl antagonists.

We used rational and combinatorial approaches to engineer an Axl "decoy receptor" that binds to the Gas6 ligand with ultra-high affinity and inhibits its

function (Kariolis et al. 2014). Upon fusion to an antibody fragment crystallizable (Fc) domain, the engineered Axl decoy receptor binds Gas6 with an affinity of ~400 femtomolar, placing it among the tightest protein-protein interactions found in nature. Crystallographic analysis of the ligand/receptor interaction, carried out in collaboration with Irimpan Mathews (SLAC National Accelerator Laboratory), showed that mutations in Axl induced structural alterations that resulted in increased Gas6/Axl binding (Figure 1).

The engineered Axl decoy receptor effectively sequestered Gas6, allowing complete abrogation of Axl signaling. Moreover, Gas6 binding affinity was critical and correlative with the ability of decoy receptors to effectively inhibit metastasis and disease progression. The engineered Axl decoy receptor inhibited up to 90 percent of metastatic nodules in two murine models of ovarian cancer compared to wild-type Axl (~50 percent inhibition), with virtually no toxic side effects (Kariolis et al. 2014).

INSPIRATION FROM NATURE TO DEVELOP A NOVEL CLASS OF TUMOR TARGETING AGENTS

A major obstacle to the development of therapeutics that target the brain is the presence of the blood-brain barrier, which prevents foreign particles and

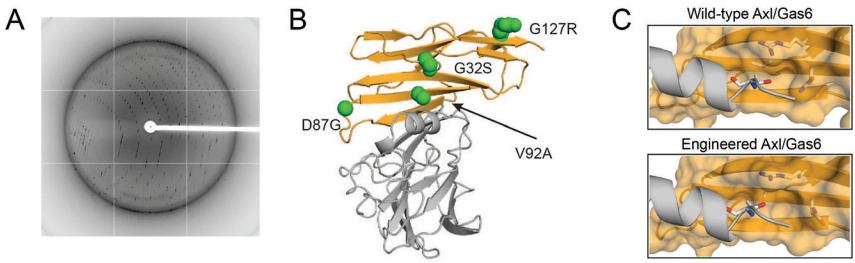

FIGURE 1 Structural analysis of an engineered protein therapeutic elucidates how it binds its target. (A) Protein crystal mounted in an x-ray beam line. Black spots represent data showing the organization of atoms in the crystal. Data from I. Mathews, SLAC National Accelerator Laboratory. (B) Structure of the engineered Axl receptor decoy in complex with Gas6 ligand: Gas6 LG1 domain (grey) and engineered Axl Ig1 domain (orange). Green spheres indicate locations of Axl mutations identified from protein engineering screen. (C) Close-up images of the wild-type Axl/Gas6 and engineered Axl/Gas6 interfaces. In the engineered version, the mutation of valine at position 92 of Axl to alanine creates a larger pocket that reinforces the structure of a key helix on Gas6. These high-resolution images provide a molecular snapshot of structural alterations in the engineered Axl/Gas6 interface that confer high affinity binding. Images B and C reprinted with permission from Kariolis et al. (2014). Figure can be viewed in color at https://www.nap.edu/catalog/23659.

molecules from entering the central nervous system. We recently demonstrated the promise of using engineered peptides, known as knottins, to target brain tumors for applications such as image-guided resection and targeted drug delivery (Ackerman et al. 2014a; Kintzing and Cochran 2016).

Knottins are unique peptides (30–50 amino acids) containing a disulfide-bonded core that confers outstanding proteolytic resistance and thermal stability (Kolmar 2009). They are found in a wide variety of plants, animals, insects, and fungi, and carry out diverse functions such as ion channel inhibition, enzyme inactivation, and antimicrobial activity (Zhu et al. 2003).

We used molecular engineering approaches to redirect a knottin found in squash seeds that normally functions as an enzyme inhibitor, to create an engineered knottin that binds tumor-associated receptors with high affinity (Kimura et al. 2009a). In collaboration with Zheng Cheng and Sanjiv Sam Gambhir (Stanford Radiology) we established these engineered peptides as a new class of molecular imaging agents for cancer (Kimura et al. 2009b) (Figure 2). We then showed that intravenous injection of an engineered knottin, conjugated to a near-infrared fluorescent dye molecule, targeted and illuminated intracranial brain tumors in animal models of medulloblastoma (collaborations with Matthew Scott, Stanford Developmental Biology, and Samuel Cheshier and Gerald Grant, Stanford Neurosurgery) (Ackerman et al. 2014b; Moore et al. 2013).

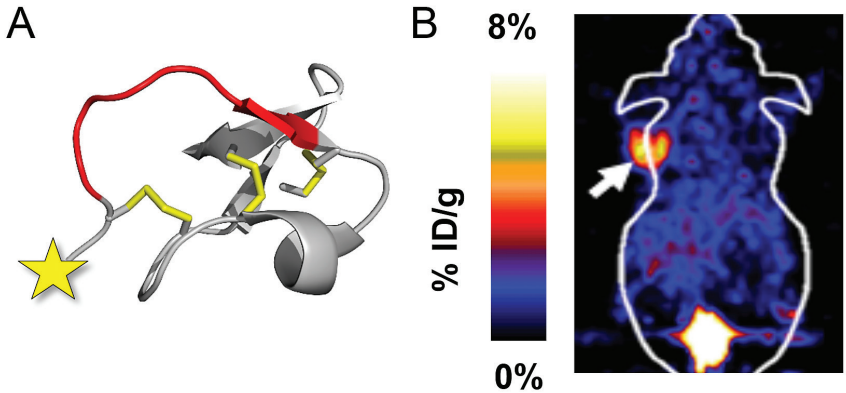

FIGURE 2 (A) 3D structural representation of a trypsin inhibitor peptide from Ecballium elaterium II (EETI-II), which was used as a starting point for engineering a tumor-targeting agent. Star indicates attachment site for molecular imaging probe. Protein Data Bank ID: 2IT7. (B) Positron emission tomography (PET) image after injection of a radiolabeled engineered knottin peptide in a mouse model of cancer. Image was acquired 1 hr postinjection. Arrow designates tumor. Scale represents percent injected dose per gram (%ID/g) and is a quantitative measure of imaging signal. Figure can be viewed in color at https://www.nap.edu/catalog/23659.

Disulfide-rich peptides, including knottins, have generated great interest as potential drug candidates as they offer the pharmacological benefits of small molecule drugs along with the target-binding affinity and specificity of protein biologics. We postulated that if we could use an engineered knottin peptide to visualize tumors, then we could also use it as a vehicle to deliver drugs to tumors, with a goal of minimizing toxic side effects of systemic chemotherapy.

In one study, carried out in collaboration with the Stanford ChEM-H Medicinal Chemistry Knowledge Center, the engineered knottin peptide was conjugated to the nucleoside analogue gemcitabine, using a variety of linker strategies, and an optimal candidate was shown to inhibit proliferation of breast, ovarian, pancreatic, and brain tumor cells in vitro (Cox et al. 2016). Notably, this peptide-drug conjugate was shown to kill cells via receptor-mediated internalization, and thus exhibited increased potency against pancreatic cells that acquired some resistance to treatment with gemcitabine alone.

In a second study, carried out in collaboration with Sutro Biopharma, Inc., the engineered peptide was fused to an antibody Fc domain and conjugated to the tubulin inhibitor monomethyl-auristatin-F. This knottin-Fc-drug conjugate was capable of inducing regression and prolonged survival in a flank glioblastoma model (Currier et al. 2016), highlighting promise for further clinical development.

CONCLUSIONS

Research and development efforts over the past few decades have culminated in a growing number of FDA-approved protein therapeutics that enable targeted treatment of cancer. In parallel, continued efforts to develop safer and more effective cancer therapeutics are being fueled by expanding knowledge of mechanisms underlying disease pathophysiology and the ability to customize proteins using a variety of engineering methods.

The case studies presented above provide examples of how our research team is using protein design and engineering to generate next-generation cancer therapeutics. Protein engineers are also using these powerful technologies to create molecular toolkits for answering a wide range of research questions in basic science, biotechnology, and biomedicine.

REFERENCES

Ackerman SE, Currier NV, Bergen JM, Cochran JR. 2014a. Cystine-knot peptides: Emerging tools for cancer imaging and therapy. Expert Review of Proteomics 11:561–572.

Ackerman SE, Wilson CM, Kahn SA, Kintzing JR, Jindal DA, Cheshier SH, Grant GA, Cochran JR. 2014b. A bioengineered peptide that localizes to and illuminates medulloblastoma: A new tool with potential for fluorescence-guided surgical resection. Cureus 6(9):e207.

Cox N, Kintzing JR, Smith M, Grant GA, Cochran JR. 2016. Integrin-targeting knottin peptide-drug conjugates are potent inhibitors of tumor cell proliferation. Angewandte Chemie International Edition 55(34):9894–9897.

Currier NV, Ackerman SE, Kintzing JR, Chen R, Filsinger Interrante M, Steiner A, Sato AK, Cochran JR. 2016. Targeted drug delivery with an integrin-binding knottin-Fc-MMAF conjugate produced by cell-free protein synthesis. Molecular Cancer Therapeutics 15(6):1291–1300.

Drewe E, Powell RJ. 2002. Clinically useful monoclonal antibodies in treatment. Journal of Clinical Pathology 55(2):81–85.

Ecker DM, Jones SD, Levine HL. 2015. The therapeutic monoclonal antibody market. MAbs 7(1):9–14.

Gustafsson A, Martuszewska D, Johansson M, Ekman C, Hafizi S, Ljungberg B, Dahlback B. 2009. Differential expression of Axl and Gas6 in renal cell carcinoma reflecting tumor advancement and survival. Clinical Cancer Research 15:4742–4749.

Hong J, Peng D, Chen Z, Sehdev V, Belkhiri A. 2013. Abl regulation by Axl promotes cisplatin resistance in esophageal cancer. Cancer Research 73:331–340.

Kariolis MS, Kapur S, Cochran JR. 2013. Beyond antibodies: Using biological principles to guide the development of next-generation protein therapeutics. Current Opinion in Biotechnology 24:1072–1077.

Kariolis MS, Miao YR, Jones DS 2nd, Kapur S, Mathews II, Giaccia AJ, Cochran JR. 2014. An engineered Axl "decoy receptor" effectively silences the Gas6-Axl signaling axis. Nature Chemical Biology 10:977–983.

Kimura RH, Levin AM, Cochran FV, Cochran JR. 2009a. Engineered cystine knot peptides that bind alphavbeta3, alphavbeta5, and alpha5beta1 integrins with low-nanomolar affinity. Proteins 77(2):359–369.

Kimura RH, Cheng Z, Gambhir SS, Cochran JR. 2009b. Engineered knottin peptides: A new class of agents for imaging integrin expression in living subjects. Cancer Research 69:2435–2442.

Kintzing JR, Cochran JR. 2016. Engineered knottin peptides as diagnostics, therapeutics, and drug delivery vehicles. Current Opinion in Chemical Biology 34:143–150.

Kolmar H. 2009. Biological diversity and therapeutic potential of natural and engineered cystine knot miniproteins. Current Opinion in Pharmacology 9(5):608–614.

Li Y, Ye X, Tan C, Hongo JA, Zha J, Liu J, Kallop D, Ludlam MJC, Pei L. 2009. Axl as a potential therapeutic target in cancer: Role of Axl in tumor growth, metastasis and angiogenesis. Oncogene 28:3442–3455.

Linger RM, Keating AK, Earp HS, Graham DK. 2008. Tam receptor tyrosine kinases: Biologic functions, signaling, and potential therapeutic targeting in human cancer. Advances in Cancer Research 100:35–83.

Moore SJ, Hayden Gephart MG, Bergen JM, Su YS, Rayburn H, Scott MP, Cochran JR. 2013. Engineered knottin peptide enables noninvasive optical imaging of intracranial medulloblastoma. Proceedings of the National Academy of Sciences 110(36):14598–14603.

Rankin EB, Fuh KC, Taylor TE, Krieg AJ, Musser M, Yuan J, Wei K, Kuo CJ, Longacre TA, Giaccia AJ. 2010. Axl is an essential factor and therapeutic target for metastatic ovarian cancer. Cancer Research 70(19):7570–7579.

Vajkoczy P, Knyazev P, Kunkel A, Capelle HH, Behrndt S, von Tengg-Kobligk H, Kiessling F, Eichelsbacher U, Essig M, Read TA, Erber R, Ullrich A. 2006. Dominant-negative inhibition of the Axl receptor tyrosine kinase suppresses brain tumor cell growth and invasion and prolongs survival. Proceedings of the National Academy of Sciences 103:5799–5804.

Walsh G. 2014. Biopharmaceutical benchmarks 2014. Nature Biotechnology 32:992–1000.

Weidle UH, Auer J, Brinkmann U, Georges G, Tiefenthaler G. 2013. The emerging role of new protein scaffold-based agents for treatment of cancer. Cancer Genomics and Proteomics 13:155–168.

Zhu S, Darbon H, Dyason K, Verdonck F, Tytgat J. 2003. Evolutionary origin of inhibitor cystine knot peptides. FASEB Journal 17:1765–1767.

Engineering Immunotherapy

Darrell J. Irvine
Massachusetts Institute of Technology

Immunotherapy aims to promote an immune response to disease. Pursued for more than 30 years as a potential treatment for cancer, it is based on the capacity of the immune system to safely distinguish healthy cells from tumor cells and to be resistant to mutational escape by tumors, and on the possibility of establishing immune memory to prevent recurrence.

THE NEW AGE OF IMMUNOTHERAPY

For many years treatments targeting the immune system showed only anecdotal efficacy in clinical trials, leading many researchers to become disillusioned with the field by the late 1990s. Yet the 1990s were a period when many critical elements of fundamental biology regulating the immune response were identified or characterized: the first tumor antigens, Toll-like receptors and related signaling pathways that govern inflammation and the immune system's ability to identify "danger," regulatory receptors that promote or block T cell activation, and specific mechanisms used by tumor cells to avoid immune destruction.

These discoveries led to a transformation in the field of immuno-oncology, which was most prominently impacted by clinical studies, in the early 2000s, of an antibody that blocks a key negative regulatory receptor on T cells, cytotoxic T lymphocyte antigen-4 (CTLA-4). Treatment of melanoma patients with this antibody enabled endogenous antitumor immune responses that led to tumor regressions in a small proportion of heavily pretreated patients with metastatic disease. About 20 percent of the patients survived more than 5 years, well beyond the expected lifespan for advanced disease (Hodi et al. 2010; Lebbé et al. 2014). This "tail of the curve" effect in overall survival reflects a dramatic change in

outcome from the best modern "targeted" therapies, where early tumor regression is generally followed by drug resistance, relapse, and death.

Following these early findings, a second class of antibodies blocking another negative regulator axis in T cells, antibodies to PD-1 on T cells (or to its ligand, PD-L1 expressed on tumor cells), showed even more dramatic effects in large clinical trials. Among patients with melanoma, renal cell carcinoma, and lung cancer, 30–50 percent showed tumor regressions (Topalian et al. 2012). These drugs, although acting by distinct mechanisms, are collectively referred to as "checkpoint blockade" therapies, as they disrupt regulatory checkpoints that restrain the immune response to cancer.

In parallel to these advances, a second type of immunotherapy approach has been developed: adoptive cell therapy (ACT), based on the transfer of autologous tumor-specific T cells into patients. In ACT, T cells are isolated from the peripheral blood or from tumor biopsies, cultured with the patient's own tumor cells to identify tumor-reactive clones, and then expanded to large numbers for reinfusion into the patient (Rosenberg and Restifo 2015). The creation ex vivo of an army of tumor-specific T cells has been shown to elicit objective tumor regressions when combined with appropriate adjuvant treatments that promote the functionality of the transferred T cells (e.g., administration of adjuvant drugs such as interleukin-2).

Other strategies genetically modify T cells for patients by introducing a synthetic T cell receptor (chimeric antigen receptor, or CAR) that allows any T cell to become a tumor-specific T cell. These have shown particular promise in treating certain leukemias: greater than 75 percent of patients have experienced complete remissions (Maude et al. 2014).

Thus, in the space of a few short years the field of cancer immunotherapy has been revolutionized in the clinic, from a peripheral approach notorious for high toxicity and low efficacy, to a frontline treatment with the prospect of eliciting durable responses—and perhaps cures—in some patients.

ROLE OF ENGINEERING IN THE FUTURE OF CANCER IMMUNOTHERAPY

Immunology has advanced by embracing new technologies, from the early days of monoclonal antibody technology to the recent inventions of powerful mass spectrometry–based cellular analysis tools.

The field has also recently attracted the attention of a growing number of interdisciplinary scientists, who bring to bear a unique mindset and new approaches to problems in immunology and immunotherapy. Some of these techniques are rooted in engineering, leading to exciting advances in basic science and new approaches to vaccines and immunotherapies.

Engineers excel at creating model systems that break complex problems into manageable hurdles, and at drawing on applied chemistry, physics, and mathemat-

ics to create new technologies that solve practical problems. Engineering contributions to the evolution of cancer immunotherapy can be illustrated by recent examples in the areas of cancer vaccines and ACT. These by no means represent all the areas where engineers are actively working on cancer immunotherapy, but rather are two representative examples.

Enhancing Cancer Vaccines

As mentioned, checkpoint blockade with anti-CTLA-4 or anti-PD-1 has elicited objective tumor regressions in a small proportion of patients. This incomplete response rate has motivated a strong interest in finding additional treatments that can be combined with these drugs to expand the responding population.

Because these drugs act to enhance T cell responses against tumors, one obvious strategy is to combine checkpoint blockade with therapeutic cancer vaccines, for patients whose spontaneous T cell responses to tumors may be too weak to be rescued by checkpoint blockade alone. To this end, a renewed interest in cancer vaccines has been kindled in both preclinical and clinical studies. However, cancer vaccines to date have generally been perceived as a failure, both because of their lack of objective responses in patients and their inability to elicit the kind of robust T cell priming that is believed to be necessary for tumor regression (i.e., T cell responses more like those to live infectious agents).

How can the efficacy of cancer vaccines be improved?

Engineered Antigens

Vaccines are generally based on the delivery of antigens (the protein, peptide, or polysaccharide target of the immune response) together with inflammatory cues that stimulate the immune system to respond to the antigens.

One of the simplest approaches that has been most extensively explored in the clinic is the use of peptide antigens combined with adjuvants as T cell–focused vaccines. But short peptides injected in vivo have several significant limitations: they are quickly degraded, they largely flush into the bloodstream rather than traffick to lymphatics and lymph nodes, and they can be presented by any nucleated cell to T cells. The latter phenomenon, in which T cells are stimulated by random tissue cells rather than professional antigen-presenting cells (APCs) in lymph nodes, leads to tolerance or deletion of tumor-specific cells.

One way to deal with all of these challenges at once is to conjugate so-called "long" peptide antigens (that can be presented only by professional APCs) to an albumin-binding lipid tail through a water-soluble polymer spacer. Albumin constitutively traffics from blood to lymph, and, thus linking antigens to an albumin-binding lipid "tail," redirects these molecules efficiently to lymph nodes instead of the bloodstream after parenteral injection. In addition, the polymer/lipid linkage

protects the peptide from degradation. A similar strategy can be used to create "albumin hitchhiking" adjuvants.

These simple chemical modifications lead to 15- to 30-fold increases in vaccine accumulation in lymph nodes, both enhancing the safety of the vaccine and dramatically increasing vaccine potency (Liu et al. 2014).

Regenerative Scaffolds

Engineers have also used methods developed in the regenerative medicine field to create implantable vaccine "centers" that coordinate multiple steps in an anticancer vaccine response. A common strategy in regenerative medicine is to create biodegradable polymeric scaffolds as artificial environments that can protect and nurture therapeutic cells on implantation in vivo.

Mooney, Dranoff, and colleagues demonstrated that a similar approach can be used to regulate the response to a vaccine (Ali et al. 2009). By loading polymeric sponges with tumor antigens, chemoattractants for APCs, and adjuvants, they coordinated a 3-step process of (1) APC attraction to the implanted scaffold, (2) uptake of antigen and adjuvant by the APCs, and (3) migration of the now activated APCs to draining lymph nodes, where they could initiate a potent antitumor immune response. This approach is currently being tested in a phase I clinical trial.

Thus chemistry and biomaterials approaches offer a number of ways to create enhanced cancer vaccines.

Engineering Adoptive Cell Therapy

As noted above, adoptive transfer of tumor antigen-specific T cells is one of the two classes of immunotherapies to demonstrate significant durable responses in the clinic so far, but strategies to improve this treatment for elimination of solid tumors are still sought.

Engineers are contributing to the evolution of ACT treatments through the application of synthetic biology principles for the creation of novel genetically engineered T cells. Recently, for example, bioengineers have generated completely artificial ligand-receptor-transcription factor systems, which enable the introduction of a synthetic receptor and transcription factor pair into T cells to enable T cell recognition of a tumor-associated ligand to be transduced into production of an arbitrary biological response (Morsut et al. 2016; Roybal et al. 2016).

Another strategy introduces synthetic fragmented antigen receptors that are activated only when a small molecule drug is present, to allow precise control over the activity of therapeutic T cells in vivo (Wu et al. 2015). These are only a few representative examples of a rapidly moving and exciting area of research.

A third strategy chemically engineers T cells using an approach from the nanotechnology and drug delivery communities to "adjuvant" T cells with supporting drugs, such as cytokines that promote T cell function and proliferation.

One promising approach is to attach drug-releasing nanoparticles directly to the plasma membrane of ACT T cells so that the modified cells carry supporting drugs on their surface wherever they home in vivo. This approach has been shown to greatly augment the expansion and antitumor activity of T cells when used to deliver supporting cytokines to the donor cells (Stephan et al. 2010). This basic demonstration also opens the potential to target supporting drugs directly to T cells in vivo, through targeted nanoparticle formulations (Zheng et al. 2013). Such studies show promise in preclinical models and are entering the early stages of translation into clinical testing.

CONCLUSIONS

Cancer therapy is being revolutionized by the first successful immunotherapy treatments. It has also created exciting new opportunities for engineers to impact the field of cancer immunotherapy, by solving challenging problems to safely enhance the immune response to tumors.

The marriage of cutting-edge tools from engineering with the latest understanding of the immune response to tumors offers the promise of further advances toward the goal of curing cancer or rendering many cancers a manageable, chronic condition.

REFERENCES

Ali OA, Huebsch N, Cao L, Dranoff G, Mooney DJ. 2009. Infection-mimicking materials to program dendritic cells in situ. Nature Materials 8:151–158.

Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, and 19 others. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. New England Journal of Medicine 363:711–723.

Lebbé C, Weber JS, Maio M, Neyns B, Harmankaya K, Hamid O, O'Day SJ, Konto C, Cykowski L, McHenry MB, Wolchok JD. 2014. Survival follow-up and ipilimumab retreatment of patients with advanced melanoma who received ipilimumab in prior phase II studies. Annals of Oncology 25:2277–2284.

Liu H, Moynihan KD, Zheng Y, Szeto GL, Li AV, Huang B, Van Egeren DS, Park C, Irvine DJ. 2014. Structure-based programming of lymph-node targeting in molecular vaccines. Nature 507:519–522.

Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, Chew A, Gonzalez VE, Zheng Z, Lacey SF, and 9 others. 2014. Chimeric antigen receptor T cells for sustained remissions in leukemia. New England Journal of Medicine 371:1507–1517.

Morsut L, Roybal KT, Xiong X, Gordley RM, Coyle SM, Thomson M, Lim WA. 2016. Engineering customized cell sensing and response behaviors using synthetic notch receptors. Cell 164:780–791.

Rosenberg SA, Restifo NP. 2015. Adoptive cell transfer as personalized immunotherapy for human cancer. Science 348:62–68.

Roybal KT, Rupp LJ, Morsut L, Walker WJ, McNally KA, Park JS, Lim WA. 2016. Precision tumor recognition by T cells with combinatorial antigen-sensing circuits. Cell 164:770–779.

Stephan MT, Moon JJ, Um SH, Bershteyn A, Irvine DJ. 2010. Therapeutic cell engineering with surface-conjugated synthetic nanoparticles. Nature Medicine 16:1035–1041.

Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, Powderly JD, Carvajal RD, Sosman JA, Atkins MB, and 20 others. 2012. Safety, activity, and immune correlates of anti-pd-1 antibody in cancer. New England Journal of Medicine 366:2443–2454.

Wu CY, Roybal KT, Puchner EM, Onuffer J, Lim WA. 2015. Remote control of therapeutic T cells through a small molecule-gated chimeric receptor. Science 350(6258):aab4077.

Zheng Y, Stephan MT, Gai SA, Abraham W, Shearer A, Irvine DJ. 2013. In vivo targeting of adoptively transferred T-cells with antibody- and cytokine-conjugated liposomes. Journal of Controlled Release 172:426–435.

APPENDIXES

Contributors

Kevin Alexander is vice president and western regional manager at Hazen and Sawyer. His research focuses on the planning, design, and construction of water, wastewater, and water reclamation facilities. He applies advanced treatment technology for high water recovery applications including reverse osmosis and other membrane and nonmembrane-based technologies.

Lars Blackmore is the principal rocket landing engineer at SpaceX, where he focuses on precision landing for space vehicles. Most recently, his team designed the algorithms and operations for the SpaceX Falcon 9 Reusable rocket. Previously he worked on precision Mars landing and autonomous air and sea vehicles. He specializes in using convex optimization to solve previously intractable onboard guidance and control problems.

Robert Braun is dean of engineering and applied science at the University of Colorado Boulder. His research has focused on planetary entry systems, Mars landing systems design, the Mars Sample Return Project, and multidisciplinary optimization.

Julie Champion is an associate professor of chemical and biomolecular engineering at the Georgia Institute of Technology. Her research is on design and fabrication of therapeutic biomaterials self-assembled from engineered proteins for applications in vaccines and in treating cancer and inflammation. Her group seeks to understand and control the interactions between these materials and cells or proteins through manipulation of molecular and physical biomaterial properties.

Amy Childress is professor and director of the environmental engineering program at the University of Southern California. Her research interests center on membrane contactor processes for innovative solutions to water treatment challenges, pressure-driven membrane processes as industry standards for desalination and water reuse, membrane bioreactor technology, colloidal and interfacial aspects of membrane processes, and salinity gradient energy production.

Jennifer Cochran is an associate professor of bioengineering at Stanford University, where she develops and uses new technology to engineer proteins for biotechnology and medical applications. Current interests include engineered biomolecules for use as diagnostic agents, cancer therapeutics, materials for tissue regeneration, and research tools for probing complex biological systems at the molecular scale.

Kayvon Fatahalian is an assistant professor of computer science at Carnegie Mellon University, where he researches the design of high-performance systems for real-time rendering and the analysis and mining of visual data at scale.

Kristen Grauman is an associate professor of computer science at the University of Texas at Austin. Her research in computer vision and machine learning focuses on visual search and recognition. Current interests include egocentric vision, language and vision, interactive segmentation, activity recognition, and video summarization.

Warren Hunt is a research scientist at Oculus Research, where he works on graphics, particularly image syntheses and display interface for virtual reality devices.

Darrell Irvine is a professor of materials science and engineering and biological engineering at the Massachusetts Institute of Technology. His research focuses on the application of engineering tools to problems in cellular immunology, the development of new materials for vaccine and drug delivery, and vaccine development for HIV and immunotherapy of cancer.

DeShawn Jackson is a production enhancement business analyst at Halliburton. Her research interests include production enhancement and subsurface insight utilizing fracture mapping and reservoir monitoring services such as microseismic monitoring, surface and downhole microdeformation analysis, distributed temperature and acoustic sensing with fiber optic monitoring, and integrated far-field and near-wellbore sensors.

Sangbae Kim is an associate professor of mechanical engineering at the Massachusetts Institute of Technology, where he focuses on bioinspired robotics, biomechanics of locomotion, and printable robots.

Brian Kirby is an associate professor of mechanical and aerospace engineering at Cornell University. His research is on microfluidics devices for biochemical analysis with applications to counterbioterrorism, environmental monitoring, medical devices, and biology. His scientific expertise includes coupling of chemistry, fluid mechanics, and electrodynamics in micro- and nanofabricated systems as well as tissue-engineered scaffolds.

Manish Kumar is an assistant professor of chemical engineering at Pennsylvania State University. His group develops materials that combine the exquisite specificity and functionality of biological molecules with the physical toughness and engineering ability of polymers. Their favorite system is combining water channel proteins (aquaporins) with block copolymers to develop next-generation desalination membranes. The group also works on innovative ideas to improve the sustainability of conventional desalination membranes.

David Lentink is an assistant professor of mechanical engineering at Stanford University, where he studies biological flight as an inspiration for engineering design. His comparative biological flight research ranges from maple seeds and insects to birds such as swifts, lovebirds, and hummingbirds. His group applies mechanical research of dynamically morphing wings, vortex dynamics, and fluid-structure interaction to robot designs that fly in complex environments in realistic atmospheric conditions.

David Luebke is vice president of research at NVIDIA, where he focuses on a variety of computer graphics research topics, especially virtual and augmented reality, real-time rendering, ray tracing, display technology, and GPU computing.

Baoxia Mi is an assistant professor of civil and environmental engineering at the University of California, Berkeley. Her research interests include membrane separation, transport, and interfacial phenomena; physicochemical processes; drinking water purification and wastewater reuse; desalination; environmental nanotechnology; and innovative applications of membrane technology for renewable energy generation, public health protection, and hygiene and sanitation improvement for underdeveloped and disaster-ridden regions.

John Orcutt is a distinguished professor of geophysics at the Cecil H. and Ida M. Green Institute for Geophysics and Planetary Physics at Scripps Institution of

Oceanography, University of California, San Diego (UCSD) and a member of the executive committee and distinguished scholar at the San Diego Supercomputer Center at UCSD. His major areas of research are marine seismology applied to both crustal and mantle structure, particularly seismic tomography; long-term ocean observations and wireless networking related to observations; theoretical seismology; and applications of seismology to monitoring of nuclear tests.

John Owens is a professor of electrical and computer engineering at the University of California, Davis. His research focuses on commodity parallel computing/GPU computing with a concentration in fundamental GPU parallel primitives (data structures and algorithms), interactive and offline computer graphics and high-performance computing applications, and multi-GPU computing. He has a recent interest in programmability of GPUs and GPUs in data centers.

Derek Paley is the Willis H. Young Jr. Associate Professor of Aerospace Engineering Education at the University of Maryland, where he conducts research in dynamics and control, including cooperative control of autonomous vehicles, adaptive sampling with mobile networks, and spatial modeling of biological groups.

Marco Pavone is an assistant professor of aeronautics and astronautics at Stanford University. His research interests are in the development of methodologies for the analysis, design, and control of autonomous systems, with an emphasis on robotic networks, autonomous aerospace vehicles, and mobility platforms for extreme planetary environments.

Cynthia Reinhart-King is an associate professor of biomedical engineering at Cornell University, where she focuses on understanding the mechanisms that drive tissue formation and tissue disruption during diseases such as atherosclerosis and cancer. Specifically, her team employs multidisciplinary methodologies involving principles from cell biology, biophysics, biomaterials, and biomechanics to study how physical and chemical cues within the extracellular environment drive fundamental cellular processes including cell-matrix adhesion, cell-cell adhesion, and cell migration.

Abhishek Roy is a senior research scientist in energy and water solutions at The Dow Chemical Company, where he uses the principles of fundamental science to create solutions for sustainable water management. His research centers on reverse osmosis technology, polymer synthesis and transport modeling of proton exchange membranes, high-temperature liquid chromatography, polyolefins, and industrial coatings.

Christopher Stafford is a materials science and engineering research chemist at the National Institute of Standards and Technology. His research is on advanced measurements of thin-film composite membranes, namely the active layer that is extremely thin and fragile, using measurements that include x-ray and neutron scattering, vibrational spectroscopy, and surface analysis.

Peter Tessier is the Richard Baruch M.D. Career Development Associate Professor of Chemical and Biological Engineering at Rensselaer Polytechnic Institute. His research is on designing and optimizing a class of large therapeutic proteins (antibodies) that holds great potential for detecting and treating human disorders ranging from cancer to Alzheimer's disease. His team designs antibodies with high binding affinity and solubility through key fundamental breakthroughs.

Gordon Wetzstein is an assistant professor of electrical engineering at Stanford University, where his group focuses on advancing imaging, microscopy, and display systems. His research at the intersection of computer graphics, machine vision, optics, scientific computing, and perception has a wide range of applications in next-generation consumer electronics, scientific imaging, human-computer interaction, and remote sensing.

Participants

Andrew Adamczyk
Senior Principal Research Engineer
Corporate Research and Development
Air Products and Chemicals

Gagan Aggarwal
Senior Staff Research Scientist
Google Research

Naoko Akiya
Platform Leader
Packaging and Specialty Plastics
Materials Science
Dow Chemical Company

Kevin Alexander**
Vice President
Hazen and Sawyer

Saleema Amershi
Researcher
Microsoft Research

Ines Azevedo
Associate Professor
Department of Engineering and Public
Policy
Carnegie Mellon University

Rajan Bhattacharyya
Senior Research Engineer
Information and Systems Sciences
Laboratory
HRL Laboratories

Lars Blackmore**
Principal Rocket Landing Engineer
Guidance Navigation and Control
SpaceX

Robert Braun*
Dean of Engineering and Applied
Science
University of Colorado Boulder

*Organizing Committee
**Speaker

Heidi Buck
Director, Battlespace Exploitation of Mixed Reality Lab
SPAWAR Automated Imagery Analysis Group
Space and Naval Space Warfare Systems Center

Qing Cao
Research Staff Member
Physical Science
IBM

Rebecca Carrier
Associate Professor
Department of Chemical Engineering
Northeastern University

Julie Champion*
Associate Professor
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Ranveer Chandra
Principal Researcher
Microsoft

Amy Childress*
Professor
Sonny Astani Department of Civil and Environmental Engineering
University of Southern California

Ian Clark
Systems Engineer
Jet Propulsion Laboratory

Jennifer Cochran**
Associate Professor
Department of Bioengineering
Stanford University

Anne Dailly
Staff Researcher
Chemical and Materials Systems Laboratory
General Motors

Seth Darling
Scientist and Co-Lead for Argonne's Water Initiative
Nanoscience and Technology Division
Argonne National Laboratory

Neil Dasgupta
Assistant Professor
Department of Mechanical Engineering
University of Michigan

Alexander Dunn
Assistant Professor
Department of Chemical Engineering
Stanford University

Katherine Dykes
Senior Engineer
National Wind Technology Center
National Renewable Energy Laboratory

Hoda Eldardiry
Manager, Machine Learning Group
Interaction and Analytics Laboratory
PARC, A Xerox Company

Jeffrey Erickson
Engineer
Center for Biomolecular Science and Engineering
Naval Research Laboratory

Rebecca Erikson
Senior Staff Engineer
Systems Engineering and Integration Division
National Security Directorate
Pacific Northwest National Laboratory

Jo Etter
Product Development Engineer
Software, Electronic and Mechanical Systems Laboratory
3M Company

Kayvon Fatahalian**
Assistant Professor
Department of Computer Science
Carnegie Mellon University

Stacey Finley
Assistant Professor
Department of Biomedical Engineering
University of Southern California

Jason Furtney
Senior Engineer
Research and Consulting
Itasca Consulting Group

Kelly Gardner
CEO
Zephyrus Biosciences, Inc.

Phanindra Garimella
Director, Dynamic Systems and Controls
Cummins

Andrew Goodwin
Assistant Professor
Department of Chemical and Biological Engineering
University of Colorado Boulder

Kristen Grauman**
Associate Professor
Department of Computer Science
University of Texas at Austin

Zhen Gu
Associate Professor
Joint Department of Biomedical Engineering
University of North Carolina at Chapel Hill and North Carolina State University

Jin-Oh Hahn
Assistant Professor
Department of Mechanical Engineering
University of Maryland

Brendan Harley
Associate Professor and Schaefer Faculty Scholar
Department of Chemical and Biomolecular Engineering
University of Illinois at Urbana-Champaign

Amy Herhold
Director, Physics and Mathematical Sciences Laboratory
Corporate Strategic Research
ExxonMobil

Warren Hunt**
Research Scientist
Oculus Research

Mahmoud Hussein
Associate Professor and H. Joseph Smead Faculty Fellow
Department of Aerospace Engineering Sciences
University of Colorado Boulder

Margot Hutchins
Research and Development Engineer
Homeland Security and Defense Systems
Sandia National Laboratories

Darrell Irvine**
Professor
Departments of Materials Science & Engineering and Biological Engineering
Massachusetts Institute of Technology

DeShawn Jackson*
Business Analyst
Production Enhancement
Halliburton

Leah Johnson
Manager, Advanced Materials
Innovation, Technology, and Development
RTI International

Joseph Kakande
Advanced Photonics Research Technical Staff
Bell Labs, Nokia

Sung Kang
Assistant Professor
Department of Mechanical Engineering
Johns Hopkins University

Amin Karbasi
Assistant Professor
Department of Electrical Engineering and Computer Science
Yale University

Laura Kennedy
Assistant Group Leader
Applied Space Systems Group
MIT Lincoln Laboratory

Branko Kerkez
Assistant Professor
Department of Civil and Environmental Engineering
University of Michigan

Sangbae Kim**
Associate Professor
Department of Mechanical Engineering
Massachusetts Institute of Technology

Brian Kirby**
Associate Professor
Sibley School of Mechanical and Aerospace Engineering
Cornell University

Manish Kumar**
Assistant Professor
Department of Chemical Engineering
Pennsylvania State University

Pankaj Kumar
Research Engineer
Modern Control Methods and Computational Intelligence
Ford Motor Company

Mariel Lavieri
Associate Professor
Department of Industrial and Operations Engineering
University of Michigan

Daeyeon Lee
Professor
Department of Chemical and Biomolecular Engineering
University of Pennsylvania

Jennifer Leight
Assistant Professor
Department of Biomedical Engineering
Ohio State University

David Lentink**
Assistant Professor
Department of Mechanical Engineering
Stanford University

Joe Lester
Senior Engineer
Upstream Packaging Process R&D
Procter & Gamble Company

Qizhen (Katherine) Li
Associate Professor
School of Mechanical and Materials Engineering
Washington State University

Jingmei Liang
Director and Process Engineer
Dielectric System & Module
Applied Materials, Inc.

Shihong Lin
Assistant Professor
Department of Civil and Environmental Engineering
Vanderbilt University

David Luebke*
Vice President, Graphics Research
NVIDIA

Richard Lunt
Associate Professor
Department of Chemical Engineering and Material Sciences
Michigan State University

Olav Lyngberg
Scientific Fellow
Advanced Technology, Technical Operations
Johnson & Johnson

Nina Mahmoudian
Assistant Professor
Department of Mechanical Engineering – Engineering Mechanics
Michigan Technological University

Elisabeth Malsch
Vice President, Forensic Engineering
Thornton Tomasetti

Joel McDonald
Engineered Materials Product Development Leader
Dow Corning Corporation

Baoxia Mi**
Assistant Professor
Department of Civil and Environmental Engineering
University of California, Berkeley

Jeremy Munday
Assistant Professor
Institute for Research in Electronics and Applied Physics
University of Maryland

John Owens*
Professor
Department of Electrical and Computer Engineering
University of California, Davis

Corinne Packard
Assistant Professor
Department of Metallurgical and Materials Engineering
Colorado School of Mines

Derek Paley**
Willis H. Young Jr. Associate Professor of Aerospace Engineering Education
Department of Aerospace Engineering and Institute for Systems Research
University of Maryland

Marco Pavone*
Assistant Professor
Department of Aeronautics and Astronautics
Stanford University

Devesh Ranjan
Associate Professor and J. Erskine Love Jr. Faculty Fellow
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Roderick Reber
Senior Engineer
Technical Polymers R&D
Arkema Inc.

Cynthia Reinhart-King**
Associate Professor
Department of Biomedical Engineering
Cornell University

Julian Rimoli
Assistant Professor
Department of Aerospace Engineering
Georgia Institute of Technology

Reuben Rohrschneider
Senior Engineer
Mission Systems Engineering
Ball Aerospace and Technologies

Julio Romero Aguero
Vice President, Strategy and Business Innovation
Quanta Technology

Abhishek Roy*
Senior Research Scientist
Energy and Water Solutions
The Dow Chemical Company

Andrea Schmidt
Research Engineer
Lawrence Livermore National Laboratory

Kelly Schultz
P. C. Rossin Assistant Professor
Department of Chemical and Biomolecular Engineering
Lehigh University

Christine Scotti
CFD Technology Leader
Medical Products Division
W.L. Gore and Associates, Inc.

Meredith Sellers
Managing Engineer
Materials and Corrosion Engineering
Exponent

Debbie Senesky
Assistant Professor
Department of Aeronautics and Astronautics
Stanford University

Robert Shepherd
Assistant Professor
Department of Mechanical and Aerospace Engineering
Cornell University

Alexander Simpson
Strategic Operations Leader
Executive Engineering: Aviation Engineering Division
GE Aviation

Thomas Simpson
Chemical Engineering Consultant
DuPont

Christopher Stafford**
Research Chemist
Materials Science and Engineering Division
National Institute of Standards and Technology

Leia Stirling
Charles Stark Draper Assistant Professor
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology

Amit Surana
Principal Research Scientist
Systems Department
United Technologies

Zoya Svitkina
Senior Software Engineer
Google Inc.

Joseph-Paul Swinski
Computer Engineer
Flight Software Systems Branch
NASA Goddard Space Flight Center

Ilias Tagkopoulos
Associate Professor
Computer Science and Genome Center
University of California, Davis

Peter Tessier*
Richard Baruch M.D. Career Development Associate Professor
Department of Chemical and Biological Engineering
Rensselaer Polytechnic Institute

Marija Trcka
Technology Sourcing Specialist
Innovation Business Development
United Technologies Corp.

Bao Truong
Lead Nuclear System Engineer
Nuclear Plant Design and Safety
Terrapower

Vassilis Varveropoulos
Drilling Systems Architecture Manager
Schlumberger

Jean Vettel
Neuroscientist and Science Lead
Human Research Engineering Directorate
Army Research Laboratory

Laura Waller
Assistant Professor
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

Xuan (Kelly) Wei
Director, R&D
Restorative Therapy Group
Medtronic

Gordon Wetzstein**
Assistant Professor
Department of Electrical Engineering
Stanford University

Edward Whalen
Associate Technical Fellow and Flight Engineering Manager
Boeing Company

Gregory Whiting
Staff, Rapid Evaluation
Google[X]

Hans (Chris) Woithe
Communication Systems Software Technical Staff
Bell Labs, Nokia

Paul Wooster
Manager
Guidance, Navigation, and Control
SpaceX

Chang Yuan
Senior Research Manager
Apple

Michail Zavlanos
Assistant Professor
Department of Mechanical Engineering and Materials Science
Duke University

Dinner Speaker

John Orcutt
Distinguished Professor of Geophysics
University of California, San Diego

Guests

William (Bill) Hayden
Vice President
The Grainger Foundation

Sohi Rastegar
Senior Advisor
Emerging Frontiers and Multidisciplinary Activities
Directorate for Engineering
National Science Foundation

National Academy of Engineering

C. D. Mote, Jr.
President

Alton D. Romig, Jr.
Executive Officer

Janet Hunziker
Senior Program Officer

Sherri Hunter
Program Coordinator

Program

NATIONAL ACADEMY OF ENGINEERING

2016 US Frontiers of Engineering
September 19-21, 2016

Chair: Robert Braun, Georgia Institute of Technology*

PIXELS AT SCALE: HIGH-PERFORMANCE COMPUTER GRAPHICS AND VISION

Organizers:
David Luebke, NVIDIA Research, and
John Owens, University of California, Davis

Computational Near-Eye Displays: Engineering the Interface to the Digital World
Gordon Wetzstein, Stanford University

Frontiers in Virtual Reality Headsets
Warren Hunt, Oculus Research

First-Person Computational Vision
Kristen Grauman, University of Texas at Austin

A Quintillion Live Pixels: The Challenge of Continuously Interpreting and Organizing the World's Visual Information
Kayvon Fatahalian, Carnegie Mellon University

***

*Currently at the University of Colorado Boulder.

EXTREME ENGINEERING: EXTREME AUTONOMY IN SPACE, AIR, LAND, AND UNDER WATER

Organizers:
DeShawn Jackson, Halliburton, and Marco Pavone, Stanford University

Autonomous Precision Landing of Space Rockets
Lars Blackmore, SpaceX

Avian Flight as an Inspiration for Drone Design
David Lentink, Stanford University

MIT Cheetah: New Design Paradigm for Mobile Robots
Sangbae Kim, Massachusetts Institute of Technology

Autonomy Under Water:
Ocean Sampling by Autonomous Underwater Vehicles
Derek Paley, University of Maryland

***

WATER DESALINATION AND PURIFICATION

Organizers:
Amy Childress, University of Southern California, and
Abhishek Roy, The Dow Chemical Company

Water Desalination: History, Advances, and Challenges
Manish Kumar, Pennsylvania State University

Scalable Manufacturing of Layer-by-Layer
Membranes for Water Purification
Christopher Stafford, National Institute of Standards and Technology

New Materials for Emerging Desalination Technologies
Baoxia Mi, University of California, Berkeley

High-Recovery Desalination and Water Treatment
Kevin Alexander, Hazen and Sawyer

***

TECHNOLOGIES FOR UNDERSTANDING AND TREATING CANCER

Organizers:
Julie Champion, Georgia Institute of Technology, and
Peter Tessier, Rensselaer Polytechnic Institute

How Cancer Cells Go Awry:
The Role of Mechanobiology in Cancer Research
Cynthia Reinhart-King, Cornell University

Advances in Detecting Rare Cancer Cells
Brian Kirby, Cornell University

Engineered Proteins for Visualizing and Treating Cancer
Jennifer Cochran, Stanford University

Engineering Immunotherapy
Darrell Irvine, Massachusetts Institute of Technology