Dædalus

Journal of the American Academy of Arts & Sciences

Fall 2014

From Atoms to the Stars

He Helium
Ne Neon
Ar Argon
Kr Krypton
Xe Xenon
Rn Radon
Uuo Ununoctium unknown
F Fluorine
Cl Chlorine
Br Bromine
I Iodine
At Astatine
Uus Ununseptium unknown
O Oxygen
S Sulfur
Se Selenium
Te Tellurium
Po Polonium
Lv Livermorium
N Nitrogen
P Phosphorus
As Arsenic
Sb Antimony
Bi Bismuth
Uup Ununpentium unknown
C Carbon
Si Silicon
Ge Germanium
Sn Tin
Pb Lead
Fl
Lu Lutetium
Lr Lawrencium
Yb Ytterbium
No Nobelium
Tm Thulium
Md Mendelevium
Cr
V Vanadium
Ti Titanium
Sc Scandium
Ac Actinium
Be Beryllium
Mg Magnesium
Ca Calcium
Sr Strontium
H Hydrogen
Li Lithium
Na Sodium
K Potassium
Rb Rubidium
Cs Cesium

Inside front cover : “Life seen as a miracle of organized atoms.”
Artwork by Adam Hurwitz. © American Academy of Arts & Sciences.

Jerrold Meinwald and Jeremiah Ostriker, Guest Editors

Phyllis S. Bendell, Managing Editor and Director of Publications

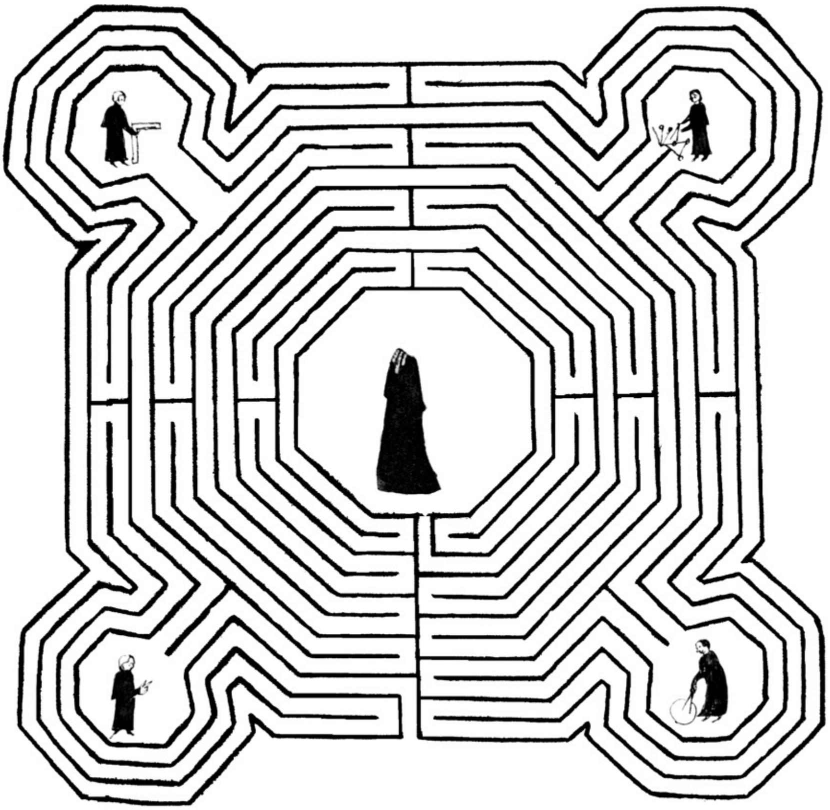
Peter Walton, Assistant Editor

Emma Goldhammer, Senior Editorial Assistant

Dædalus is designed by Alvin Eisenman.

Dædalus

Journal of the American Academy of Arts & Sciences



The pavement labyrinth once in the nave of Reims Cathedral (1240), in a drawing, with figures of the architects, by Jacques Cellier (c. 1550 – 1620)

Dædalus was founded in 1955 and established as a quarterly in 1958. The journal's namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. Its emblem, a maze seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its nearly five thousand elected members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Fall 2014
Issued as Volume 143, Number 4

Editorial offices: *Dædalus*, American Academy of Arts & Sciences, 136 Irving Street, Cambridge MA 02138. Phone: 617 576 5085. Fax: 617 576 5088. Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299

Dædalus publishes by invitation only and assumes no responsibility for unsolicited manuscripts. The views expressed are those of the author of each article, and not necessarily of the American Academy of Arts & Sciences.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, One Rogers Street, Cambridge MA 02142-1209, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from The MIT Press. Subscription and address changes should be addressed to MIT Press Journals Customer Service, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2889; U.S./Canada 800 207 8354. Fax: 617 577 1545. Email: journals-cs@mit.edu.

Printed in the United States of America by Cadmus Professional Communications, Science Press Division, 300 West Chestnut Street, Ephrata PA 17522.

Newsstand distribution by Ingram Periodicals Inc., 18 Ingram Blvd., La Vergne TN 37086.

Postmaster: Send address changes to *Dædalus*, One Rogers Street, Cambridge MA 02142-1209. Periodicals postage paid at Boston MA and at additional mailing offices.

Subscription rates: Electronic only for nonmember individuals – $47; institutions – $129. Canadians add 5% GST. Print and electronic for nonmember individuals – $52; institutions – $144. Canadians add 5% GST. Outside the United States and Canada add $23 for postage and handling. Prices subject to change without notice.

Institutional subscriptions are on a volume-year basis. All other subscriptions begin with the next available issue.

Single issues: $13 for individuals; $35 for institutions. Outside the United States and Canada add $6 per issue for postage and handling. Prices subject to change without notice.

Claims for missing issues will be honored free of charge if made within three months of the publication date of the issue. Claims may be submitted to journals-cs@mit.edu. Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising and mailing-list inquiries may be addressed to Marketing Department, MIT Press Journals, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Introduction

Jerrold Meinwald

Why "From Atoms to the Stars"? In the Summer 2012 issue of *Dædalus*, entitled "Science in the 21st Century," May Berenbaum and I sought to provide representative accounts of recent progress in the natural sciences. But it turned out that two areas of the physical sciences – astronomy and chemistry – cried out for more extensive attention than we were then able to provide. Consequently, Jeremiah Ostriker and I recruited a group of outstanding astronomers and chemists to write a set of essays to complement this earlier issue. Each of the new essays in this volume discusses important scientific developments in astronomy and chemistry in specific areas of study to which the authors themselves have made major contributions.

Philosophers, alchemists, and subsequently chemists have examined the properties and transformations of matter in all its diversity for over two millennia. The pace of progress of these studies (and, in fact, in all areas of science) picked up markedly toward the end of the eighteenth century, and has been increasing rapidly ever since. It was not until twenty years after the first performance of Stravinsky's *The Rite of Spring* (and, I was shocked to realize, during my own lifetime!) that it became clear, with James Chadwick's discovery of the neutron in 1932, that all ordinary matter is made up simply of protons, neutrons, and electrons. While protons and neutrons were at first believed to be the fundamental particles making up atomic nuclei, they have since the 1960s been best understood as *composite subatomic*

JERROLD MEINWALD, a Fellow of the American Academy since 1970, is the Goldwin Smith Professor of Chemistry Emeritus at Cornell University. His research has contributed to a wide range of chemical and chemical biological subjects, including organic photochemistry, reaction mechanisms, the synthesis of chiral inhalation anesthetics, natural product chemistry, and chemical ecology. His publications include the edited volumes *Chemical Ecology: The Chemistry of Biotic Interaction* (with Thomas Eisner, 1995) and *Science and the Educated American: A Core Component of Liberal Education* (with John G. Hildebrand, 2010). He is Secretary of the American Academy and Cochair of the Academy's Committee on Studies and Publications.

doi:10.1162/DAED_e_00300

particles, each made up of three inseparable quarks. Protons (which carry a single positive charge) consist of two *up quarks* and one *down quark*. Neutrons (electrically neutral) comprise one up quark and two down quarks. Interestingly, this revolutionary structural insight into the nature of matter has had no impact at all on our understanding of chemistry.

The simplest atom, hydrogen (H), which is not only the most abundant form of ordinary matter in the observable universe but also the most abundant atom in our own bodies, consists of a single nuclear proton and a single planetary electron. The addition of one or two neutrons to the proton yields the hydrogen isotopes deuterium and tritium, respectively. The combination of two nuclear protons and two neutrons, along with two planetary electrons, produces an atom of helium (He). With the exception of tiny amounts of lithium (Li), whose nucleus contains three protons, these are the sole types of atom produced as a consequence of the Big Bang some 13.8 billion years ago. From a chemist's viewpoint, there are some extremely important differences between hydrogen and helium atoms. Hydrogen atoms are able to bond to many other types of atoms to form stable molecules such as elemental hydrogen (H_2), water (H_2O), ammonia (NH_3), methane (CH_4), and literally millions of other "organic" compounds (all of which also contain carbon). Helium atoms, in contrast, prefer to remain alone.

It was not until the mid-nineteenth century (1869) that Dmitri Mendeleev taught us that all the known elements, when listed according to increasing atomic number (the number of protons in the nucleus), could be arranged into a "periodic table." His table revealed that the fewer than one hundred naturally occurring elements fall into periodically recurring groups (such as noble gases and halogens), a finding that enabled him to predict (correctly) the existence of unknown "missing" elements that remained for future research to discover. Our understanding of how, where, and when all the elements with an atomic number greater than two (quaintly referred to as "metals" within the astronomical community) were produced is much more recent and still somewhat incomplete. Anna Frebel's account in this volume of the origin of these elements following the Big Bang is a fascinating story that is much less well known (even among chemists) than it deserves to be. Her essay (among the astronomy contributions) provides an ideal introduction to the very existence of chemistry and of life itself.

Christopher Cummins responded to our invitation to write an essay on inorganic chemistry by examining the chemistry of a single element: phosphorus (P). In his exploration of phosphorous, we learn that pyrophoric (spontaneously combustible) elemental white phosphorus consists of discrete P_4 molecules in which each phosphorus atom occupies the vertex of a tetrahedron, the simplest of the five Platonic solids. (By a striking coincidence, Plato tells us that Timaeus considered the "element" *fire* to be composed of tetrahedral particles!)

Cummins's research on how the synthesis of phosphorus-containing compounds might be greatly simplified exemplifies theoretical and experimental chemical thinking at its best. From a consideration of the complex genesis and low abundance of phosphorus in the universe (where it is relatively rare) and in living organisms (where it occurs in concentrations much higher than it does in our solar system), we move on to accounts of its vital importance in agriculture and industry.

Tracing the passage of phosphorus from the soil into plants, then into animals, and finally into the sea illuminates some seriously underappreciated ecological concerns. The sad tale of the rise and fall of the chemically innovative Canadian research project that aimed at improving agricultural phosphorus use by creating a breed of pig, the *Enviropig*, able to digest plant-derived phosphorus compounds in its diet better than any previously existing breed of pig, brings this essay to a close. It may come as a surprise that the study of a single element reaches into such a wide range of human concerns, spanning agriculture, industry, the health sciences, ecology, and even the social sciences. Readers of Cummins's essay will be rewarded not only with insights into some beautiful science, but also with extensive material for chemistry-based cocktail party conversation.

The borders between the classical scientific disciplines are rapidly disappearing, but continuing in the more or less traditional fields of *inorganic* and *physical chemistry*, John Meurig Thomas has given us an intriguing essay on chemical catalysis, embedded in a wide-ranging examination of the importance of unpredictability and chance within and beyond chemistry. His vision of the "chemist" leans in the direction of what used to be termed "natural philosopher." He provides a refreshing view of the world of chemical research, with an emphasis on the importance of entirely unanticipated discoveries and unforeseen practical applications. His case studies serve to remind us of the remarkable value of curiosity-driven research. There is an important message here for society at large with respect to shaping the most productive science policy.

The discipline of chemistry has quietly undergone an absolutely remarkable transformation (or perhaps expansion would be a better term) over the last half-century. This development has manifested itself in part through the examination of biology as a molecular science. With our increasing understanding of the chemistry of proteins, nucleic acids, and the myriad "small molecules" that serve as molecular messengers throughout nature, dramatic and even unimaginable improvements in the practices of medicine (including psychiatry) and agriculture are certain to play a prominent role in the twenty-first century. In another direction, with the successful synthesis first of the simplest organic molecules (urea, ethyl alcohol, vanillin) and then of many of the much more complex structures (cholesterol, vitamin B-12, insulin) far behind us, can the construction of synthetic viruses and even living cells be far off?

Another major opportunity for research that is occupying the attention of many contemporary chemists is the development of *materials science*, an area discussed in Fred Wudl's essay. As a result of many advances in physics, we know vastly more than we did only a few decades ago about the way atomic and molecular interactions influence the macroscopic properties of all sorts of materials. We know why some materials are brittle, flexible, good electrical conductors, good electrical insulators, magnets, or light emitters or absorbers. As the physics of all these phenomena is better understood, it becomes possible for chemists to design and produce new materials from which everything from "improved" fabrics to computers, airplanes, and televisions can be made. Quite remarkably, the element carbon plays a central role in the design of many of the novel materials with desirable properties, such as "self-healing" plastic (vitrimers) or solar photovoltaic cells. Wudl's essay outlines how this area of chemistry evolved and what we may expect from it in the decades to come.

Chaitan Khosla has focused his essay on the field that has become known as *chemical biology*. Perhaps influenced by the ancient Greek aphorism "know thyself," he places particular emphasis on the roles that chemistry plays in understanding (as well as improving) the lives of *Homo sapiens*. Chemistry lies at the heart of much medical research, from the development of noninvasive imaging techniques such as MRI and PET scans to the discovery of new molecular targets that may serve as the basis for the design of much-needed, novel anti-infective agents to help battle malaria, Lyme disease, SARS, and many other threats to human health. In some areas, the chemistry and biology relevant to health is already fairly well understood, and "translation" from theory to application can be expected to proceed smoothly. In others, such as the chemical/biological understanding of brain functioning or the control of development, basic research remains an essential precursor to human applications. Khosla's essay illuminates a field the chemical basis of which is not yet widely enough appreciated.

Chemistry is an experimental science. Some of its complexity derives from the fact that it deals typically with huge numbers of molecules at a time; after all, an ounce of water contains about 10^{24} H_2O molecules, roughly equal to the estimated number of stars in the observable universe. Nevertheless, the enormous power of contemporary computers, combined with fundamental physical insights provided by the development of quantum mechanics, has resulted in the birth of computational chemistry, which provides both explanations and predictions of chemical properties and behavior. K. N. Houk and Peng Liu describe examples of the power of computational chemistry in predicting the products of chemical reactions, in understanding the course of newly discovered catalytic reactions, and even in designing synthetic enzymes capable of catalyzing reactions for which no natural enzymes exist. Although the power of computational chemistry is now apparent, the science is still in its infancy.

The world of computational chemistry is a far cry from the chemistry labs of our youth, with their litmus paper, Bunsen burners, and distilling flasks. The pungent odors of bromine or nitrobenzene, the beauty of deep-purple potassium permanganate crystals, the brilliance of burning magnesium, the eerie blue glow of luminol treated with hydrogen peroxide (experiences that have attracted generations of young students to chemistry in the past) are absent from this new world. They are replaced by the less sensual but nevertheless deeply satisfying insights that only the computer can give! There can be no doubt that much of the sort of chemical research that is now being carried out in the conventional laboratory with actual chemicals will be done within the next few decades faster, cheaper, and more safely by computational chemists sitting in their offices.

In summation, what we have here is a five-course chemical tasting menu. It would have been possible to choose five entirely different aspects of chemistry that would have given an equally appropriate account of the rapid advance of this lively discipline. In many menus, some of the courses or wine-pairing descriptions use unfamiliar terminology. Understandably, we are inclined to avoid incomprehensible or unpronounceable items. Describing chemistry presents an analogous challenge; one of the great problems in writing about science is how to eliminate jargon. But this difficulty can be readily taken care of these days simply by googling the obscure terms (Wikipedia also offers highly informative accounts of everything from the Platonic solids, neutrons, the periodic table, and phosphogypsum to the Enviropig and PET scans). Whether chemistry always intrigued you, or whether it was your worst subject in high school, I hope you will find yourself enjoying what our dedicated authors have to say. *Bon appétit*!

Phosphorus: From the Stars to Land & Sea

Christopher C. Cummins

Abstract: The chemistry of the element phosphorus offers a window into the diverse field of inorganic chemistry. Fundamental investigations into some simple molecules containing phosphorus reveal much about the ramifications of this element's position in the periodic table and that of its neighbors. Additionally, there are many phosphorus compounds of commercial importance, and the industry surrounding this element resides at a crucial nexus of natural resource stewardship, technology, and modern agriculture. Questions about our sources of phosphorus and the applications for which we deploy it raise the provocative issue of the human role in the ongoing depletion of phosphorus deposits, as well as the transfer of phosphorus from the land into the seas.

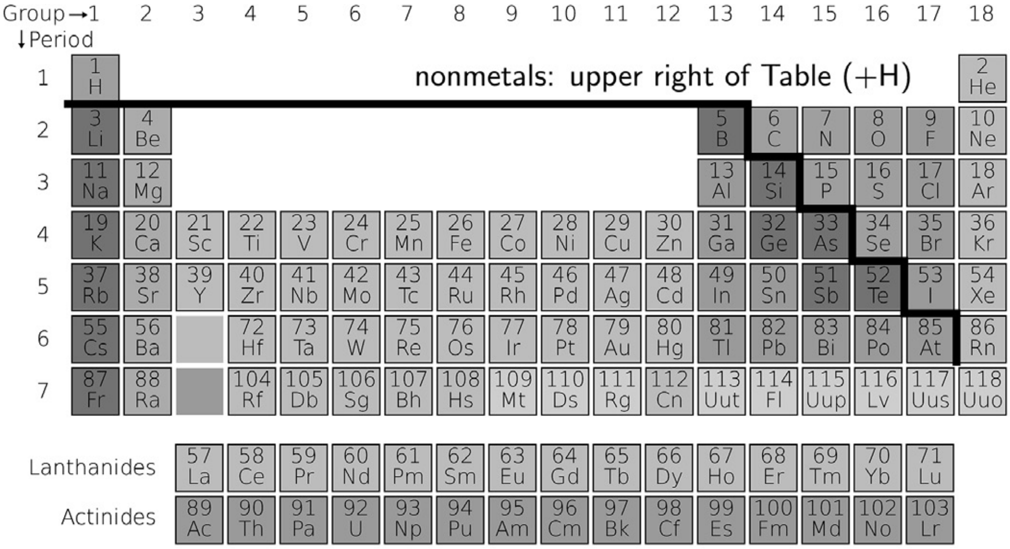
Inorganic chemistry can be defined as "the chemistry of all the elements of the periodic table,"[1] but as such, the field is impossibly broad, encompassing everything from organic chemistry to materials science and enzymology. One way to gain insight into and appreciate the rapidly moving and diverse field of inorganic chemistry is to view the science from the perspective of the elements themselves, since they are the basic ingredients for assembling molecules or materials – and indeed, all matter, living or inanimate. Although phosphorus may be less celebrated than carbon or hydrogen, it joins those elements (along with nitrogen, oxygen, and sulfur) to constitute the six "biogenic elements" (those needed in large quantities to make living organisms; see Figure 1).[2] Let us take a look at some of the issues that arise in inorganic chemistry from the perspective of phosphorus, illustrating in the process the notion that each element has its own story to tell.

Many phosphorus-containing chemical compounds are commercially valuable and have interesting or important applications.[3] Lithium hexafluorophosphate, for example, is the electrolyte in common

CHRISTOPHER C. CUMMINS, a Fellow of the American Academy since 2008, is Professor of Chemistry at the Massachusetts Institute of Technology. His research focuses on innovating new methods of inorganic synthesis, as well as the synthesis of new simple substances. His work has recently appeared in *Inorganic Chemistry*, *Science*, *Journal of the American Chemical Society*, and *Chemical Science*, among other journals.

doi:10.1162/DAED_a_00301

Figure 1
Periodic Table with Nonmetals, Including Biogenic Elements, Above the Stair-Step Line



Group→ ↓Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba		72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra		104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
Lanthanides			57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
Actinides			89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

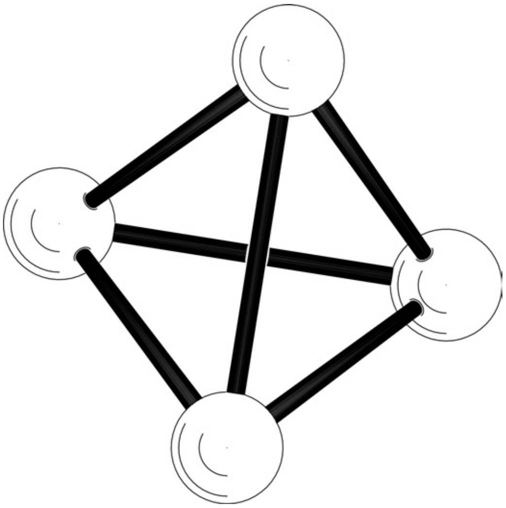
Biogenic elements are H, C, N, O, P, S in the "nonmetals" region of the periodic table, indicated by the heavy line. Source: Adapted from a graphic found on http://www.openclipart.org.

lithium-ion batteries, which are used in consumer electronics (such as laptops) and automotive applications.[4] So how is it made? The synthesis route begins with the white form of elemental phosphorus, a simple molecular form of the element consisting of tetrahedral P_4 molecules (Figure 2).[5] White phosphorus is combined with elemental chlorine in order to bring the phosphorus to the correct oxidation state (+5), and then, in a second step, chloride is replaced by fluoride.

This process is also frequently used to synthesize many organo-phosphorus compounds that are important components of catalysts used in the chemical industry.[6] In these applications, again, white phosphorus is first oxidized using chlorine, and then the chloride provides the basis for the formation of carbon-phosphorus bonds.[7] But notably, elemental chlorine is hazardous to use and ship, and environmental groups have called for an outright ban on it.[8] So why use chlorine to oxidize phosphorus if chlorine is not even present in the products, such as lithium hexafluorophosphate, that are the target of synthesis? These industry standard processes suggest there is room for improvement: if manufacturers eliminated the use of chlorine in the synthesis of important phosphorus compounds in which chlorine is absent, both hazards and waste would be significantly reduced.

Because our research has shown that it is indeed possible to derive organo-phosphorus compounds directly from white phosphorus, this is an opportunity for inorganic chemistry to improve the safety and efficiency of the manufacturing process. In one advance, we showed that phosphorus-carbon bonds can be generated by using white phosphorus together with a source of organic radicals.[9] Each of the six phosphorus-phosphorus bonds present in a molecule of white phosphorus absorbs two organic radicals in the process of being broken; each P_4 tetrahedron is broken

Figure 2
Tetrahedral Arrangement of Atoms in a P_4 (White Phosphorus) Molecule



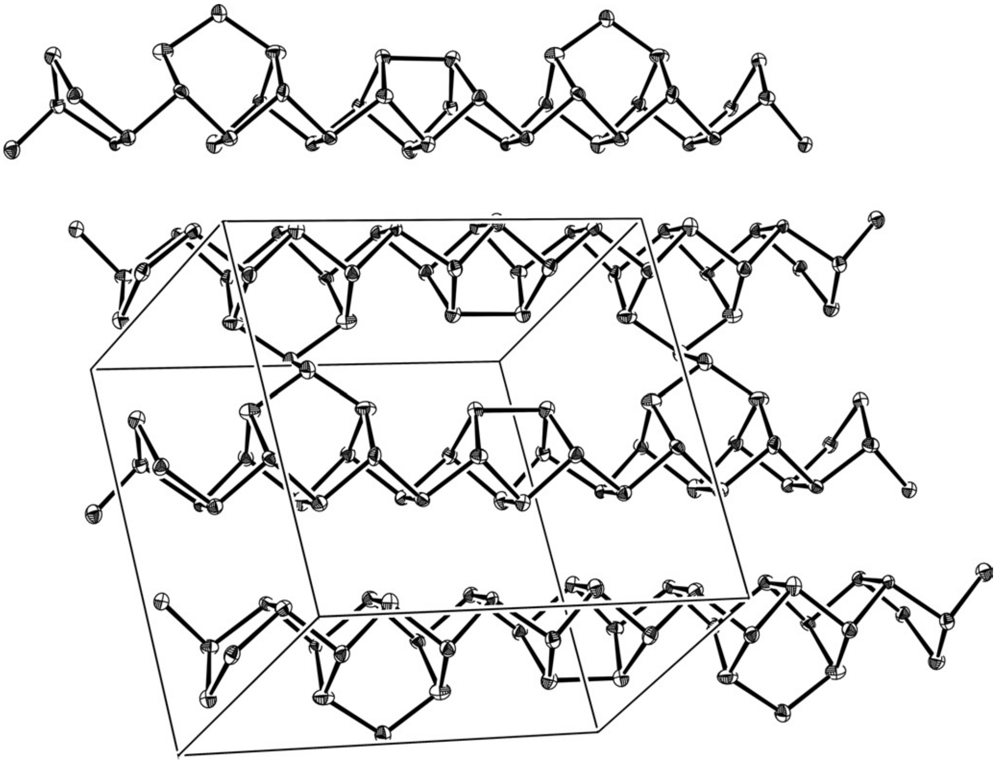
Source: Generated by the author using the PLATON program. See A. L. Spek, "Single-Crystal Structure Validation with the Program PLATON," *Journal of Applied Crystallography* 36 (2003): 7 – 13.

completely apart, and each phosphorus atom becomes incorporated into a freshly formed organo-phosphorus compound.

Our method for developing this new process was derived from basic inquiries into phosphorus's relationship to the elements neighboring it on the periodic table. Phosphorus is immediately beneath nitrogen on the periodic table, suggesting that these elements should have some similarities in their chemical properties. Then why, we wondered, was it the case that, while Earth's atmosphere consists mainly of triply-bonded N_2 molecules, a similar diatomic molecular form of phosphorus is neither prevalent nor even particularly stable?[10] Part of the answer is that nitrogen is unusual because the stability of its multiple bond far exceeds that of the sum of an equivalent number (three) of its single bonds. So the only stable form of elemental nitrogen is the diatomic molecular form floating innocuously about in the atmosphere we breathe; in contrast, phosphorus (like its diagonal relative, carbon)[11] exists in a wide variety of structural arrangements, all of which are networks exclusively based upon phosphorus-phosphorus single bonds, three for every phosphorus node. The variant known as red phosphorus, for example, has cages of phosphorus atoms connected into linear tubes (see Figure 3),[12] which in turn are cross-linked together to form a polymeric network.

Knowing this, we were inspired to ask: can we design and synthesize a molecule that would be prone to a fragmentation reaction wherein one of the fragments produced would be the diatomic molecule P_2? If we could, we would have the opportunity to study the properties and chemical characteristics of an all-phosphorus molecule structurally analogous to the main constituent of Earth's atmosphere. In our first attempt to produce it, the selected design incorporated a feature patterned after the reaction used to inflate an automobile airbag in the event of a collision, a process that rapidly generates nitrogen gas from a solid precursor. Our target molecule em-

Figure 3
Arrangement of Atoms in One of the Representative Structural Forms of Red Phosphorus

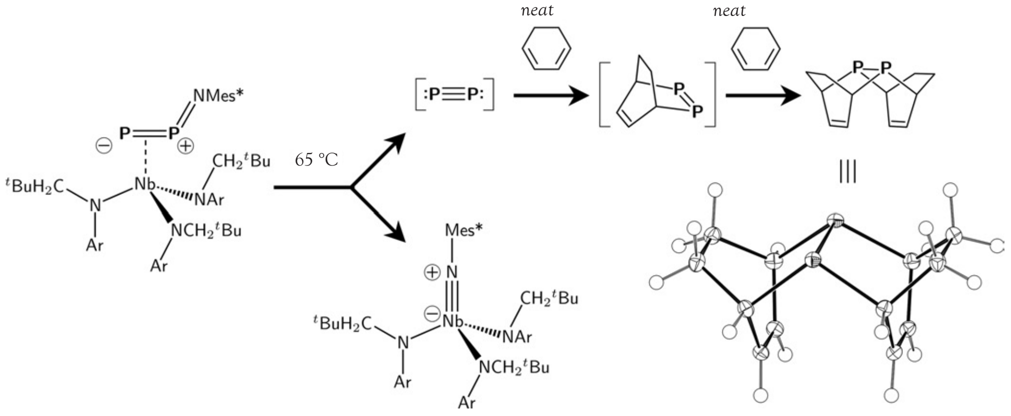


The box encloses one crystallographic unit cell. Source: Generated by the author using crystallographic coordinates from M. Ruck et al., "Fibrous Red Phosphorus," *Angewandte Chemie International Edition* 44 (2005), doi:10.1002/anie.200503017.

bedded a diphosphorus moiety into the stabilizing environment of a niobium complex (niobium is a transition metal; it forms complexes by arranging sets of molecules or ions – called ligands – around itself), from which it could be released by a stimulus of mild heating.[13] Carrying out the fragmentation reaction in the presence of other molecules permitted the mapping of the reactivity patterns of diatomic phosphorus. One important result was the discovery that P_2 easily undergoes addition to unsaturated organic molecules, such as 1,3-cyclohexadiene (see Figure 4).

If diatomic molecular phosphorus is indeed capable of direct combination with organic molecules, then the means of its generation should not matter. Could there be a way to access the P_2 molecule by starting from a stable form of the element, rather than from an exotic niobium complex? We found the suggestion in a lightly cited 1937 paper that the photochemical conversion of white phosphorus into the red form of the element may occur with P_2 as the key intermediary, which is initially generated and subsequently polymerizes.[14] We found by experiment (see Figure 5) that the addition of methyl isoprene to a solution of white phosphorus during irradiation both inhibits the production of red phosphorus *and* yields molecules in the same class of organo-phosphorus compounds that we studied earlier in connec-

Figure 4
A Niobium Complex that Can Act as an "Eliminator" of P_2 under Thermal Fragmentation



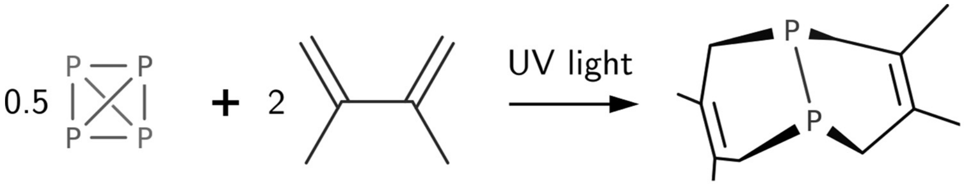
In the depicted sequence, transient P_2 (not observed) combines with two molecules of 1,3-cyclohexadiene resulting in four new P-C single bonds in the stable final product (shown both as a line drawing and in a thermal ellipsoid representation from a single-crystal X-ray diffraction analysis). Abbreviations: ^{t}Bu is tert-butyl, Ar is aryl (specifically 3,5-$Me_2C_6H_3$), and Mes* is supermesityl. Source: Adapted from material published in N. A. Piro, J. S. Figueroa, J. T. McKellar, and C. C. Cummins, "Triple-Bond Reactivity of Diphosphorus Molecules," *Science* 313 (2006), doi: 10.1126/science.1129630.

tion with niobium-mediated access to diphosphorus molecules.[15] Hence, in effect and in principle, we have shown that in certain cases the hazardous and wasteful use of chlorine in the synthesis of organophosphorus compounds can be replaced with a process relying on ultraviolet radiation.

After viewing the beautiful tetrahedral molecular form of elemental phosphorus in Figure 2, one might wonder whether this arrangement of phosphorus is unique to this particular element. Arsenic (As) lies just below phosphorus on the periodic table, separated from it by the stair-step line dividing the metals from the nonmetals (see Figure 1). Once again, the periodicity of chemical properties suggests that molecular arsenic might adopt a similar tetrahedral structure to that of phosphorus. Indeed, it does, but only in the gas phase where the molecules are well isolated from one another, or in solution at low temperature and in the dark. To generate gas-phase As_4 molecules, one heats grey arsenic (which has a layered sheet structure reminiscent of graphite or black phosphorus) to about 550 degrees C while flowing a carrier gas over it. The As_4 molecules, entrained in the carrier gas, can be led into a solvent and used for reaction chemistry before re-polymerization to grey arsenic can take place. If condensed to a solid on a cold surface, the As_4 condensate is "yellow arsenic," but it cannot be kept. Warming to room temperature or exposure to light brings about a facile return to the grey form.[16]

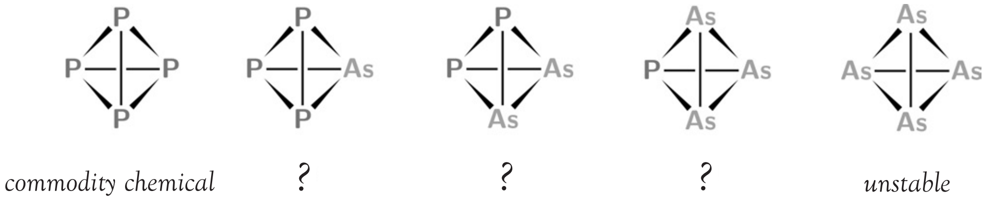
Phosphorus and arsenic lie on either side of the divide (marked on Figure 1) separating the metals from the non-metals. White phosphorus is stable enough that it can be stored as a pure liquid above its melting point of 44 degrees C and pumped into tank cars for shipping; while, conversely, samples of yellow arsenic are evanescent. We therefore wondered: would it be possible to synthesize a stable substance whose tetrahedral molecules would be composed of

Figure 5
Combination of Phosphorus with Methyl Isoprene under the Action of Ultraviolet Light



Methyl isoprene is the precursor to the original synthetic rubber. The reaction mechanism may involve P_2 as a reactive photo-generated intermediate.[17] Source: Generated by the author using data from Daniel Tofan and Christopher C. Cummins, "Photochemical Incorporation of Diphosphorus Units into Organic Molecules," *Angewandte Chemie International Edition* 49 (2010), doi:10.1002/anie.201004385.

Figure 6
From the Commodity-Chemical P_4 Molecule to the Unstable As_4 Molecule



What are the properties of substances composed of tetrahedral molecules of mixed composition? Source: Adapted from Brandi M. Cossairt and Christopher Cummins, "Properties and Reactivity Patterns of AsP_3: An Experimental and Computational Study of Group 15 Elemental Molecules," *Journal of the American Chemical Society* 131 (42) (2009), doi:10.1021/ja906294m.

a *mixture* of phosphorus and arsenic (see Figure 6)? To test this idea, we made a niobium complex carrying a P_3^{3-} unit, and combined this with a source of arsenic (3+), effectively knitting together the neutral AsP_3 molecule in a selective fashion.[18] The new substance turned out to have a waxy appearance much like that of white phosphorus, and it could be purified by sublimation, wherein the pure material is condensed onto a cold probe. Because of the volatile nature of AsP_3, we and our collaborators determined its properties by a variety of techniques, including electron diffraction, microwave spectroscopy, and photoelectron spectroscopy.[19] Obtaining gas-phase property data on a simple molecule containing a heavy element (arsenic) provides a benchmark for theorists working on the *a priori* prediction of properties; heavy elements pose the greatest challenge in this regard. The elements in the AsP_3 molecule are packaged together in a 1:3 ratio at the molecular level; and now this substance is readily available as a starting material. Substitution of a single nitrogen atom into the P_4 tetrahedron has scarcely been considered; one possibility involves stabilization inside a recently discovered spherical B_{80} molecule that is analogous to Buckminsterfullerene (C_{60}).[20]

To ask why diatomic phosphorus is neither stable nor prevalent is really to ask a larger question: why is elemental phosphorus not found on Earth as a pure substance, uncombined with other chemical elements? It is because elemental phosphorus is especially prone to oxidation, a process encouraged by Earth's atmosphere at this point in history. Elements that form very stable oxides (such as aluminum, phosphorus, and silicon) are not found in uncombined form on our planet unless they can be formed by biological or geological processes taking place under anaerobic conditions (as in the case of volcanic sulfur, or carbon in the form of coal and diamond). If we cannot obtain phosphorus in pure form directly by digging it out of the ground, where do we get it?

Phosphate rock (also known by its mineral name apatite) is essentially the bones and teeth of ancient marine organisms formed into concentrated deposits where long-evaporated seas once stood.[21] It is extracted through strip mining and forms the basis for the phosphorus fine chemicals industry. One of the principal methods for white phosphorus[22] production is the "thermal process," which involves use of an electric arc furnace, carbon in the form of coke as a reducing agent, and silica to absorb the oxide ions liberated in the heating process.[23] The elemental phosphorus is thus extracted from the rock in what is essentially an expensive purification process. Note that most phosphorus-containing commercial chemicals contain phosphorus in the +5 oxidation state: the same as is found in phosphate rock when it is dug out of the ground. The typical purification process reduces phosphorus's oxidation state from +5 to zero; however, when it is converted to other chemicals, chlorine is often used to return the phosphorus to its highest oxidation state (zero back to +5). (This method of making white phosphorus is, in fact, reminiscent of the one used by the alchemist Hennig Brand, who made phosphorus the thirteenth element to be obtained in pure form.[24] In search of the philosopher's stone, the alchemist collected great quantities of human urine, which he concentrated to a paste and subjected to reductive distillation.)

Phosphate rock is not only the basis for the fine chemicals industry of phosphorus; it is also the starting point for the (much larger) phosphorus side of the fertilizer industry. The "wet process" of purification uses sulfuric acid to generate phosphoric acid from phosphate rock, after which it can be made into critical fertilizers such as monoammonium phosphate, or MAP. Around 1940, the human population of our planet began to rise more rapidly than it had previously (since my birth in 1966 the population has doubled).[25] This critical rise in population growth coincided with two important developments in the fertilizer industry: the worldwide commercial deployment of the Haber-Bosch ammonia synthesis (whereby ammonia for agricultural applications is obtained by direct combination of the elements hydrogen and nitrogen); and the large-scale mining of phosphate rock deposits, mainly for fertilizer applications. Prior to the mid-twentieth century, humankind had been largely limited to locally available nutrients for crop production. Now, ammonia can be had in essentially limitless supply by combining the atmosphere's inexhaustible supply of nitrogen with hydrogen (which is currently derived from natural gas by steam reforming). Can phosphorus keep up?

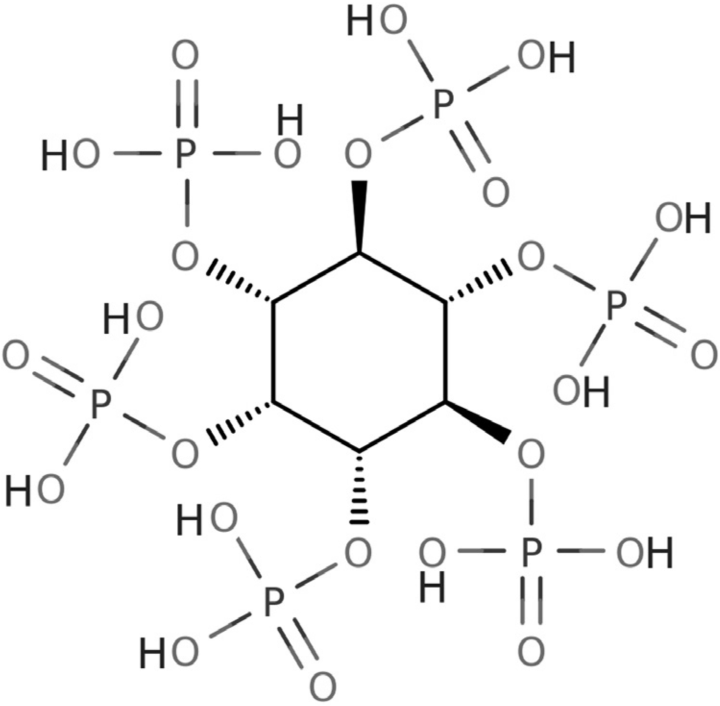
Stars are the element factories.[26] They consist mainly of our universe's lightest and most abundant elements: hydrogen and helium. Red giants are more evolved stars with an onion-like layered structure; the most abundant metallic elements, iron and nickel, make up their core, and layers of progressively lighter elements surround them, moving outward to the surface. Elements heavier than iron and nickel are formed by neutron capture when a massive star explodes in a supernova, and these (including the precious gold sought by the alchemist) are of minimal cosmic abundance. It is one of the peculiarities of nuclear physics that nuclei of odd atomic number (odd *Z*) are generally less stable and less abundant than those of even *Z*. The only stable isotope of phosphorus is ^{31}P ($Z = 15$), and the ^{31}P nucleus is the product of an extremely improbable sequence of nuclear reactions (the final reaction in the sequence converts ^{31}Si into ^{31}P by proton capture), only taking place during an explosive neon burning phase in the core of massive, hot stars.[27] Accordingly, the cosmic abundance of phosphorus is lower – by orders of magnitude – than that of the other five biogenic elements. Indeed, to quote astrobiologist Douglas Whittet: "The only biogenic element present in the human body (and in biological tissue generally) at a concentration substantially above its solar abundance is P. If one were to attempt to place an upper limit on the total biomass present in the Universe at large, on the basis of cosmic abundances, then the critical element would be phosphorus."[28] This is in keeping with the observation that, in many of the ecosystems on Earth, phosphorus is life-limiting. This means that the addition of phosphorus (usually in the form of phosphate) will bring about an abrupt bloom of life, since the absence of phosphorus was all that was holding it back.

Our land reserves of phosphorus are finite. And given the ongoing depletion of phosphate rock reserves, it is natural to ask what is left, where it is, and how long it will last. The U.S. Geological survey indicates that roughly three-quarters of the available reserves are concentrated in Morocco and Western Sahara.[29] Mining locations in Florida and Idaho contain the most significant amount of phosphate rock in the United States, but these constitute a small percentage of global reserves. And Central Florida's mines have been largely exhausted, leaving behind a legacy of radioactive phosphogypsum stacks and collapsing sinkholes.[30] The term "peak phosphorus" is now used with reference to the point in time when phosphate rock production (mining) will inevitably begin to taper off.[31] Current estimates place peak phosphorus some time later in the twenty-first century.

Off the coast of Brittany, France, there are sometimes blooms of marine algae vast enough to be visible from space.[32] Brittany is a livestock-producing region where large amounts of phosphate from feed is transferred to the ground water and ultimately to the ocean. This perfectly illustrates two consequences of the large-scale mining of phosphate rock and industrialized agricultural activity: first, we are depleting the concentrated reservoirs of this key nutrient; second, its dispersal into the world's oceans can have negative effects on marine ecosystems, chiefly by causing eutrophication through overgrowth of certain species of phytoplankton.

What can we do to mitigate the movement of phosphorus from land to sea? Efforts are being directed at optimizing the separation and recovery of phosphorus from waste water, which is an important direction.[33] In some countries (such as India and Sweden), the use of toilets that separate liquid from solid waste is being adopted; phosphate can then be recovered from urine as the crystalline mineral struvite, while solid waste is composted.[34] Pigs cannot digest plant-derived phosphate because of the phytic acid form in which plants store it (see Figure 7), so researchers at the University of Guelph in Canada developed the Enviropig.[35] This genetically engineered pig secretes the enzyme phytase in its saliva, enabling the pig to digest the plant phosphate, whereupon its excreta are phosphate-poor, leading to an improvement in waste water quality. While the meat of the Enviropig is the same as that of

Figure 7
Molecular Structure of Phytic Acid



Source: Generated by the author using the program MarvinSketch. Marvin is used for drawing, displaying, and characterizing chemical structures, substructures, and reactions. See ChemAxon, *Marvin* 6.1.3 (2014), http://www.chemaxon.com.

an unmodified pig, concerns about this creative kind of genetic engineering have effectively blocked its adoption thus far.

The chemistry of an element is a fascinating thing, and we have explored several of the issues that flow naturally from asking questions about where an element comes from, what we use it for, and how we might gain an improved understanding of it. Motivated to study phosphorus by curiosity and a desire to expand on fundamental science, we have come to appreciate the vital role played by this relatively precious element that forms the inorganic backbone of DNA, the energy currency of ATP, and the main component of bones and teeth. We have demonstrated the ability to identify ways of using this limited resource that minimize waste, but we acknowledge our limited ability to grapple with the consequences of enormous demand for phosphorus – a markedly limited resource – stemming from a rapidly rising human population. Phosphorus, therefore, is interesting not only for its chemistry but also in light of the rich texture of its larger story, only one of the many stories that emerge when we view inorganic chemistry from the perspective of a single element.

ENDNOTES

Author's Note: I am grateful for stimulating conversations with Professor Brandi M. Cossairt, Professor Anna Frebel, Professor Jerrold Meinwald, and Dr. Willem Schipper. This material is based upon work supported by the National Science Foundation under CHE-1362118.

[1] Norman Neill Greenwood and Alan Earnshaw, *Chemistry of the Elements* (Oxford: Butterworth-Heinemann, 1997).

[2] D. C. B. Whittet and J. E. Chiar, "Cosmic Evolution of the Biogenic Elements and Compounds," *The Astronomy and Astrophysics Review* 5 (1993), doi:10.1007/BF00872922.

[3] B. Elvers and U. Fritz, *Phosphorus Compounds, Inorganic to Plastics, Additives* (Weinheim, Germany: Wiley-VCH, 2011).

[4] Kazunori Ozawa, ed., *Lithium Ion Rechargeable Batteries: Materials, Technology, and New Applications* (Weinheim, Germany: Wiley-VCH, 2009).

[5] Greenwood and Earnshaw, *Chemistry of the Elements*.

[6] Paul C. J. Kamer and Piet W. N. M. van Leeuwen, *Phosphorus (III) Ligands in Homogeneous Catalysis: Design and Synthesis* (West Sussex; Chichester, U.K.: John Wiley & Sons, 2012).

[7] Ibid.

[8] R. Ayers, "The Life-Cycle of Chlorine, Part I: Chlorine Production and the Chlorine-Mercury Connection," *Journal of Industrial Ecology* 1 (1997), doi:10.1162/jiec.1997.1.1.81.

[9] Brandi M. Cossairt and Christopher C. Cummins, "Radical Synthesis of Trialkyl, Triaryl, Trisilyl and Tristannyl Phosphines from P_4," *New Journal of Chemistry* 34 (2010), doi:10.1039/c0nj00124d.

[10] W. E. Dasent, *Nonexistent Compounds: Compounds of Low Stability* (New York: M. Dekker, 1965).

[11] Keith B. Dillon, François Mathey, and John F. Nixon, *Phosphorus: The Carbon Copy: From Organophosphorus to Phospha-organic Chemistry* (New York: Wiley, 1998).

[12] M. Ruck, D. Hoppe, B. Wahl, P. Simon, Y. Wang, and G. Seifert, "Fibrous Red Phosphorus," *Angewandte Chemie International Edition* 44 (2005), doi:10.1002/anie.200503017.

[13] N. A. Piro, J. S. Figueroa, J. T. McKellar, and C. C. Cummins, "Triple-Bond Reactivity of Diphosphorus Molecules," *Science* 313 (2006), doi:10.1126/science.1129630.

[14] G. Rathenau, "Optische und photochemische versuche mit phosphor," *Physica* 4 (1937): 503–514, http://www.sciencedirect.com/science/article/B6X42-4F0H2V0-6T/2/62b38c13a477cb0dc6b01d06d9a76727.

[15] Daniel Tofan and Christopher C. Cummins, "Photochemical Incorporation of Diphosphorus Units into Organic Molecules," *Angewandte Chemie International Edition* 49 (2010), doi:10.1002/anie.201004385.

[16] A. Rodionov, R. Kalendarev, J. Eiduss, and Yu. Zhukovskii, "Polymerization of Molecular (Yellow) Arsenic," *Journal of Molecular Structure* 380 (1996), doi:10.1016/0022-2860(95)09195-5.

[17] Lee-Ping Wang, Daniel Tofan, Jiahao Chen, Troy Van Voorhis, and Christopher C. Cummins, "A Pathway to Diphosphorus from the Dissocation of Photoexcited Tetraphosphorus," *Royal Society of Chemistry Advances* 3 (2013), doi:10.1039/c3ra43940b.

[18] Brandi M. Cossairt, Mariam-Céline Diawara, and Christopher C. Cummins, "Facile Synthesis of AsP_3," *Science* 323 (2009), doi:10.1126/science.1168260.

[19] Brandi M. Cossairt, Christopher C. Cummins, Ashley R. Head, Dennis L. Lichtenberger, Raphael J. F. Berger, Stuart A. Hayes, Norbert W. Mitzel, and Gang Wu, "On the Molecular and Electronic Structures of AsP_3 and P_4," *Journal of the American Chemical Society* 132 (2010), doi:10.1021/ja102580d; and Adam M. Daly, Brandi M. Cossairt, Gavin Southwood, Spencer J. Carey, Christopher C. Cummins, and Stephen G. Kukolich, "Microwave Spectrum of Arsenic

Triphosphide," *Journal of Molecular Spectroscopy* 278 (2012): 68–71, http://www.sciencedirect.com/science/article/pii/S0022285212000410.

[20] Jules Tshishimbi Muya, Erwin Lijnen, Minh Tho Nguyen, and Arnout Ceulemans, "Encapsulation of Small Base Molecules and Tetrahedral/Cubane-Like Clusters of Group V Atoms in the Boron Buckyball: A Density Functional Theory Study," *The Journal of Physical Chemistry A* 115 (2011), doi:10.1021/jp107630q.

[21] Eric Oelkers and Eugenia Valsami-Jones, "Phosphate Mineral Reactivity and Global Sustainability," *Elements* 4 (2008): 83–87, http://elements.geoscienceworld.org/cgi/content/abstract/4/2/83.

[22] White phosphorus is a commodity chemical that is also in demand for military applications, in which setting it is valued for its incendiary effects (including the production of smoke screens) and is known as "Willy Pete." Small quantities of white phosphorus for research purposes were previously available from many chemical suppliers. After 2001, white phosphorus ceased to be available from catalog suppliers in the United States, but since red phosphorus was still available and can be converted to the white form by simple thermal depolymerization, this was not a major impediment to researchers. Now, both red and white phosphorus are Drug Enforcement Agency (DEA) List 1 controlled chemicals, which has led to difficulty in purchasing either form of the element. Why is phosphorus on this list? While there is no phosphorus whatsoever in the chemical composition of the much-abused drug methamphetamine, a popular street method for synthesizing the drug involves a combination of phosphorus and hydrogen iodide as a reducing agent for ephedrine. For the same reason, elemental iodine is also a DEA List 1 chemical. My research group was fortunate to receive a gift of white phosphorus from Thermphos (see endnote 23); after arriving from the Netherlands, however, the shipment (which was around half a kilogram) was held up in customs in New Jersey until, with the assistance of MIT's general council and a customs broker properly licensed to receive shipments of List 1 chemicals, we were able to free the shipment.

For more on the conversion of red to white phosphorus, see J. Brodkin, "Preparation of White Phosphorus from Red Phosphorus," *Journal of Chemical Education* 37 (2) (1960), doi:10.1021/ed037pA93.1. For more on the use of phosphorus in methamphetamine synthesis, see Harry F. Skinner, "Methamphetamine Synthesis via Hydriodic Acid/Red Phosphorus Reduction of Ephedrine," Forensic Science International 48 (1990): 123–134, http://www.sciencedirect.com/science/article/pii/037907389090104 7.

[23] A company called Thermphos (recently gone out of business) in the Netherlands was situated adjacent to a nuclear power plant for the cheap electricity, and took its phosphate rock shipments from Florida, now mostly mined out. Thermphos was taking a leadership role toward the exciting goal of making phosphorus from waste and thereby realizing a vision of a sustainable phosphorus industry. White phosphorus is made in the United States as the first step in the synthesis of glyphosate. See A. D. E. Grossbard, *The Herbicide Glyphosate* (London; Boston: Butterworths, 1985).

[24] John Emsley, *The 13th Element: The Sordid Tale of Murder, Fire, and Phosphorus* (Malden, Mass.: John Wiley & Sons, Inc., 2000).

[25] Jan Willem Erisman, Mark A. Sutton, James Galloway, Zbigniew Klimont, and Wilfried Winiwarter, "How a Century of Ammonia Synthesis Changed the World," *Nature Geoscience* 1 (2008), doi:10.1038/ngeo325.

[26] Anna Frebel, "Reconstructing the Cosmic Evolution of the Chemical Elements," *Dædalus* 143 (4) (2014): 71–80; and Sean G. Ryan and Andrew J. Norton, *Stellar Evolution and Nucleosynthesis* (Cambridge: Cambridge University Press, 2010).

[27] Enrique Maciá, "The Role of Phosphorus in Chemical Evolution," *Chemical Society Reviews* 34 (2005), doi:10.1039/B416855K.

[28] Whittet and Chiar, "Cosmic Evolution of the Biogenic Elements and Compounds."

[29] Tina-Simone S. Neset and Dana Cordell, "Global Phosphorus Scarcity: Identifying Synergies for a Sustainable Future," *Journal of the Science of Food and Agriculture* 92 (2012), doi:10.1002/jsfa.4650.

[30] P. Rutherford, M. Dudas, and R. Samek, "Environmental Impacts of Phosphogypsum," *Science of the Total Environment* 149 (1994), doi:10.1016/0048-9697(94)90002-7; and C. Hull and W. Burnett, "Radiochemistry of Florida Phosphogypsum," *Journal of Environmental Radioactivity* 32 (3) (1996), doi:10.1016/0265-931X(95)00061-E.

[31] Dana Cordell and Stuart White, "Peak Phosphorus: Clarifying the Key Issues of a Vigorous Debate about Long-term Phosphorus Security," *Sustainability* 3 (2011): 2027–2049, http://www.mdpi.com/2071-1050/3/10/2027.

[32] Gabriel M. Filippelli, "The Global Phosphorus Cycle: Past, Present, and Future," *Elements* 4 (2008): 89–95, http://elements.geoscienceworld.org/cgi/content/abstract/4/2/89.

[33] S. A. Parsons and J. A. Smith, "Phosphorus Removal and Recovery from Municipal Wastewaters," *Elements* 4 (2008): 109–112, http://elements.geoscienceworld.org/cgi/content/abstract/4/2/109; and Kersti Linderholm, Anne-Marie Tillman, and Jan Erik Mattsson, "Life Cycle Assessment of Phosphorus Alternatives for Swedish Agriculture," *Resources, Conservation and Recycling* 66 (2012): 27–39, http://www.sciencedirect.com/science/article/pii/S0921344912001048.

[34] Dana Cordell, Jan-Olof Drangert, and Stuart White, "The Story of Phosphorus: Global Food Security and Food for Thought," *Global Environmental Change* 19 (2009): 292–305, http://www.sciencedirect.com/science/article/pii/S095937800800099X.

[35] S. P. Golovan, R. G. Meidinger, A. Ajakaiye, M. Cottrill, M. Z. Wiederkehr, D. J. Barney, C. Plante, J. W. Pollard, M. Z. Fan, M. A. Hayes, J. Laursen, J. P. Hjorth, R. R. Hacker, J. P. Phillips, and C. W. Forsberg, "Pigs Expressing Salivary Phytase Produce Low-Phosphorus Manure," *Nature Biotechnology* 19 (2001), doi:10.1038/90788.

Foresight, Unpredictability & Chance in Chemistry & Cognate Subjects

John Meurig Thomas

Abstract: In numerous branches of natural philosophy, the ways in which major, transformative advances are achieved are often cloaked in mystery, or arrived at through a fortunate concatenation of circumstances. This theme is pursued here with the aid of some examples from my own work on catalysis (the speeding up of the attainment of chemical equilibria), as well as from the work of others. The emergence of the maser (forerunner of the laser), the development of positron emission tomography, and the creation of blood-glucose sensors for use by those suffering from type 2 diabetes are among the innovations adumbrated here. In addition to describing the unpredictable nature of much scientific discovery, I also describe areas in which new chemical technology will be especially beneficial to society. I foresee that open-structure solid catalysts are likely to transform many of the ways in which chemicals, now manufactured in an environmentally harmful manner, will be produced in the future. Also outlined is the vital need to understand and exploit photocatalysts so as to harness solar energy. Finally, I touch upon the absolute value of chemistry in the quest for beauty and truth.

JOHN MEURIG THOMAS, a Foreign Honorary Member of the American Academy since 1990, is Honorary Professor in the Department of Materials Science at the University of Cambridge and Emeritus Professor at the Davy-Faraday Laboratory, London. Formerly, he was Director of the Royal Institution of Great Britain, London, and Head of the Department of Physical Chemistry and Master (Head) of Peterhouse, Cambridge. His publications include *Principles and Practice of Heterogeneous Catalysis* (with W. J. Thomas, 2014) and *Michael Faraday and the Royal Institution: The Genius of Man and Place* (1991). He was knighted in 1991 for services to chemistry and the popularization of science.

Nearly fifty years ago, while watching Nobel laureate John Kendrew present a BBC program on the molecules of life, I was surprised to hear him remark that even expert scientists cannot usually predict what will happen in their fields more than three years in advance.[1] How could it be, I wondered, that scientific giants like him could hold such a view? Do road maps used by scientists become invalid after a mere three years? As my knowledge of advances in chemistry and adjacent fields grew, however, I began to feel that Kendrew's view is close to the truth. Before venturing into the past fortunes and future possibilities of chemistry, I will first demonstrate the veracity of Kendrew's statement with the aid of three historical examples that underline the unpredictable nature of advances in science and technology.

In 1937, President Roosevelt asked a group of expert scientists, engineers, and businessmen for advice on what developments in science and technology could likely be expected in the future, in part so that he

doi:10.1162/DAED_a_00302

could better serve his fellow Americans and foster the common good. The report correctly foresaw that agricultural science would play an ever-increasing role in the economy of the United States. It also predicted that gasoline and other useful products could be readily generated from coal.[2] But in retrospect, it is not what the experts identified as likely advances that make the report so interesting: it is the advances that were missed.[3]

Roosevelt's carefully selected experts can be forgiven for not mentioning phenomena or effects that had not yet been discovered: nuclear fission, nuclear fusion, the transistor, the laser, space satellites, and jet aircraft; or, looking further ahead, genetic engineering, genetic fingerprinting, the structure of DNA, and immunosuppressive drugs (which make organ transplantation possible). However, this collection of experts surprisingly omitted to identify antibiotics, fuel cells, fax machines, or synchrotron radiation, all of which were already known well before 1937. Perspicacious physicist members of that panel might have even identified tomography, the mathematical foundations of which had been described by mathematician Johann Radon in Leipzig in 1917. Also in 1917, Einstein published his famous paper in which Einstein coefficients (which pertain to electronic energy levels in atoms and materials) were first described. From this account, an "expert" might even have predicted the existence of masers and lasers!

My second example comes from the journal *Scientific American* (a publication to which my education owes a great deal). In 1920, as part of a special anniversary issue, it ran a series of articles authored by eminent experts on the theme of "The Future as Suggested by Developments of the Past Seventy-Five Years." Notably, an article by the selected expert on aviation claimed that, in the future, aerial travel vessels would be divided into two types: the dirigible for long-distance and transoceanic travel, and the airplane for overland routes.[4] It offered the hypothesis that "experimental work which is being carried out in producing steam drive for airplane service has given such promising results that it is quite possible a combined steam boiler and steam turbine will be in extensive use within the next few years." This never came to pass for a variety of reasons that require no elaboration here.[5]

My third example involves the discovery of the maser, the forerunner of the laser. In the 1950s at Columbia University, physicist Charles Townes became intrigued by the possibility that the population of energy levels in simple molecules could be inverted, and by what the optical consequences of such an inversion might be. When he proposed an experiment involving the inversion process, his colleague Isidor Rabi, a Nobel Prize–winning physicist, told him he was wasting his time. Other eminent scientists, including physicist Niels Bohr and mathematician John von Neumann, doubted the worthiness of the experiment. But Townes stubbornly persevered and so discovered the maser. Its logical successor, the laser, has changed our world comprehensively. (In both masers and lasers, microwave and visible light, respectively, are emitted in a polarized state, with high, adjustable intensity and a sharply defined wavelength.) In addition, Townes's work led to the discovery that nearby galaxies emit maser light, which falls upon Earth.

It is against a background of such errant and incomplete predictions on the part of unimpeachable experts that I embark on the present essay on foresight and chance in chemistry. As we shall see, some practical advances emerge through foresight and as a result of rectilinear progress in a well-trodden field; others, however, arise from unexpected sources and observations.

John Meurig Thomas

My principal areas of research as a chemist are heterogeneous catalysis and chemical electron microscopy. In the former, a solid catalyst speeds up the rate of attainment of chemical equilibrium when one, two, or more potential reactant molecules of different kinds of gases or liquids are brought together in its presence. The kind of question I seek to answer is the following. How is it that molecules impinging upon certain catalytic surfaces at velocities of typically 1600 km/h^{-1} can be converted at that surface, with high efficiency and often with spectacular selectivity, into a desired product; whereas the same substance colliding with other (inert) surfaces merely rebounds with more or less retention of translational, vibrational, and rotational energy? A good catalyst has three main characteristics: it must have high activity, meaning that it must facilitate fruitful reaction as rapidly as possible; it should be selective, meaning that it must favor just one of the various options that are thermodynamically allowed as reaction pathways, thereby yielding a desired product; and it must have longevity, meaning that it should continue to function actively and selectively for as long as possible before its catalytic performance degrades. I am preoccupied with ways of designing and synthesizing new catalysts that meet these standards.

High-school textbooks often state, erroneously, that a catalyst is an agent that does not itself undergo change. This is not quite true. All catalysts interact with the species that they transform, and in so doing, they may ultimately change their nature or become contaminated with molecular fragments that gradually diminish or eliminate their efficacy. It is sometimes possible to reactivate a "dead" catalyst by judicious chemical manipulation. For example, the solid catalysts used extensively to convert petroleum to useful products, such as gasoline or polymer precursors, gradually become poisoned by accumulated build-up of carbonaceous material at their exterior surface. By carefully burning the "poison" in air, the catalyst can be regenerated.

Although my own studies of catalysts are primarily of an academic nature, catalysis is a vitally important means of sustaining modern civilized life. One example of this is the enormous array of products derived from oil, almost always with the assistance of catalysts (see Figure 1).

Another striking reminder of how vital catalysis is to modern life is illustrated in Figure 2, which shows how the girth of a tree that grew in a Norwegian forest underwent a remarkable increase in its rate of growth once ammonia fertilizer was administered to the forest after the tree's first twenty-five years of life.

The catalyst that has long been used for the conversion of nitrogen (N_2) and hydrogen (H_2) to yield ammonia (NH_3) was discovered by the German chemist Fritz Haber in Karlsruhe in 1909. It is composed principally of metallic iron, but tinctures of potassium and alumina (Al_2O_3) are added to it to optimize its effectiveness and longevity. Many companies worldwide are able to prepare long-lived, durable, and highly active ammonia-synthesis catalysts that produce millions of metric tons of ammonia each year. Some of their iron-based catalysts last for over a decade in continuous use without significant loss of performance.

In order to devise new and better solid catalysts, it is necessary first to understand precisely how existing ones work. Often this is an intellectually and experimentally difficult task, especially when some catalysts (like the iron one for ammonia synthesis) operate under elevated temperatures and pressures. There are very few experimental techniques available for *in situ* studies of catalysts that operate (within either a thick ceramic or other refractory chamber) at nearly 400 degrees C and

Figure 1
A Selection of the Products Made from Petroleum

Gasoline
Heating oil
Tents
Crayons
Parachutes
Telephones
Enamel
Pillows
Dishes
Cameras
Anesthetics
Artificial turf
Artificial limbs
Bandages
Dentures
Model cars
Folding doors
Hair curlers
Cold cream
Movie film
Soft contact lenses
Drinking cups
Fan belts
Car enamel
Shaving cream
Ammonia
Refrigerators
Golf balls

Toothpaste
Heart valves
Candles
Trash bags
House paint
Water pipes
Hand lotion
Roller skates
Surfboards
Shampoo
Wheels
Paint rollers
Shower curtains
Guitar strings
Luggage
Aspirin
Safety glasses
Antifreeze
Awnings
Eyeglasses
Clothes
Toothbrushes
Ice chests
Footballs
Combs
CDs and DVDs
Paint brushes
Detergents

Perfumes
TV cabinets
Shag rugs
Electrician's tape
Tool racks
Car battery cases
Epoxy
Paint
Mops

Fishing rods
Lipstick
Denture adhesive
Linoleum
Ice cube trays
Synthetic rubber
Speakers
Plastic wood
Electric blankets
Glycerin

Cassettes
Dishwasher parts
Toolboxes
Transparent tape
Shoe polish
Helmets
Caulking
Petroleum jelly
CD players
Faucet washers
Antiseptics
Clothesline
Curtains
Basketballs
Soap
Vitamin capsules
Antihistamines
Purses
Food preservatives
Shoes
Dashboards
Cortisone
Deodorant
Footballs
Putty
Dyes
Pantyhose
Refrigerant

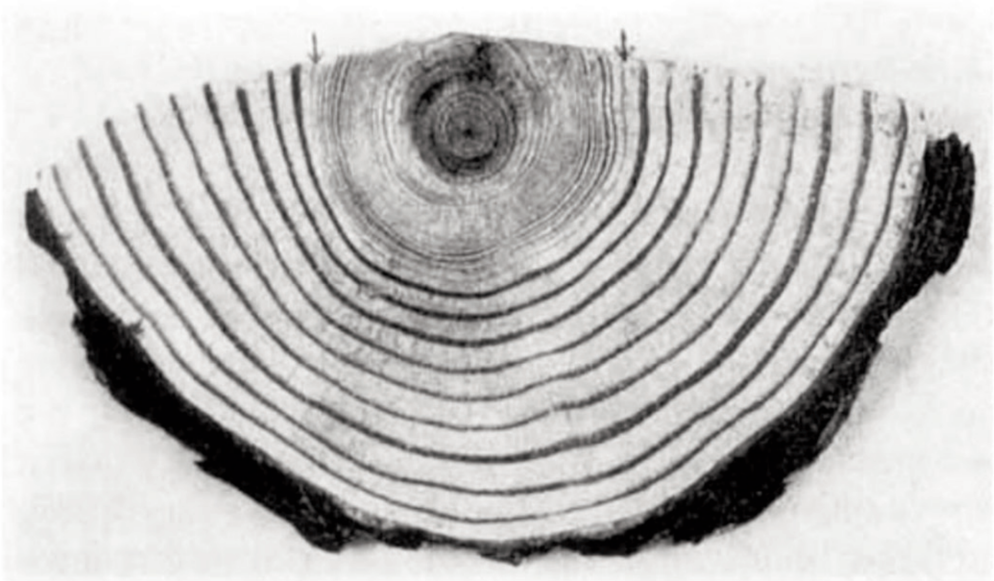
Dresses
Tires
Golf bags
Percolators
Life jackets
Rubbing alcohol
Tennis rackets
Rubber cement
Fishing boots
Vaporizers
Balloons
Sunglasses
Solvents
Diesel fuel
Motor oil
Bearing grease
Ink
Floor wax
Ballpoint pens
Football cleats
Upholstery
Sweaters
Boats
Insecticides
Bicycle tires
Sports car bodies
Nail polish
Fishing lures

Source: Prepared by Jay Keasling of the Lawrence Berkeley National Laboratory

around 200 atmospheres of pressure. Hence, it is necessary to use indirect model studies. To illustrate how difficult it is to elucidate the iron-ammonia-synthesis catalysts' modes of operation, we consider the work of Gerhard Ertl of the Fritz Haber Institute of the Max Planck Gesellschaft in Berlin. In the early 1980s, he was able after much effort to demonstrate beyond doubt the sequence of individual chemical steps that occur at the surface of the iron catalyst when it produces ammonia. Briefly, the process begins when N_2 and H_2 dissociate to yield relatively loosely bound nitrogen and hydrogen atoms at the iron surfaces. By an ingenious sequence of experiments, Ertl was able to work out the frequency of collisions and the associated energetics of these atoms at the iron surface and hence explain fully how, from transitory diatomic and triatomic fragments like NH and NH_2, gaseous NH_3 is finally (catalytically) formed. For this discovery, Ertl earned the Nobel Prize in 2007.

For the last three decades, my own work on the synthesis, characterization, and deployment of new solid catalysts has focused on open-structure solids. I first modified or synthesized clay minerals (not unlike mica) that are composed of negatively charged sheets of atoms consisting mainly of aluminum, silicon, and oxygen. I could then convert such solids into powerful solid acids – as strong as sulfuric acid – and capitalize on their consequential catalytic ac-

Figure 2
Growth Increase of a Tree Following Administration of Ammonia Fertilizer



A dendrochronological illustration of how effective fertilizers are. Helicopters sprinkled ammonia on a Norwegian forest regularly after the trees had been growing for twenty-five years. Thereafter, rapid growth occurred. Source: Image provided to the author by the late Professor Joseph Chatt, University of Sussex.

tivity. This entails populating the spaces between the sheets of the synthetic clay with positively charged hydrogen atoms (protons), which are very acidic. In this way, my colleagues at the University of Wales and I were able to produce a highly active and selective catalytic method of generating the well-known sweet fragrance ethyl acetate.[6] In fact, this advance is now utilized industrially on a massive scale in the United Kingdom to produce three hundred thousand metric tons of ethyl acetate in a one-step, solvent-free process, the attributes of which are of great importance in this age of clean technology and green chemistry.

This is just one of many examples that illustrate how catalysts can render chemical processes both clean (environmentally sustainable) and efficient (requiring less energy to generate the desired product). Because clean chemical processes will likely be demanded by legislators and the public for the indefinite future, one can safely predict that there will be a great need for catalysts that make such processes feasible. For far too long, the chemical industry, as well as researchers in our colleges and universities, have blithely used numerous chemical reagents that are noxious, toxic, potentially explosive, or otherwise hazardous. In the future, the use of powerful oxidants (such as nitric acid or potassium permanganate) or powerful acids – like oleum or hydrofluoric acid (which nowadays are widely used in chemical technology) – will almost certainly be banned.[7]

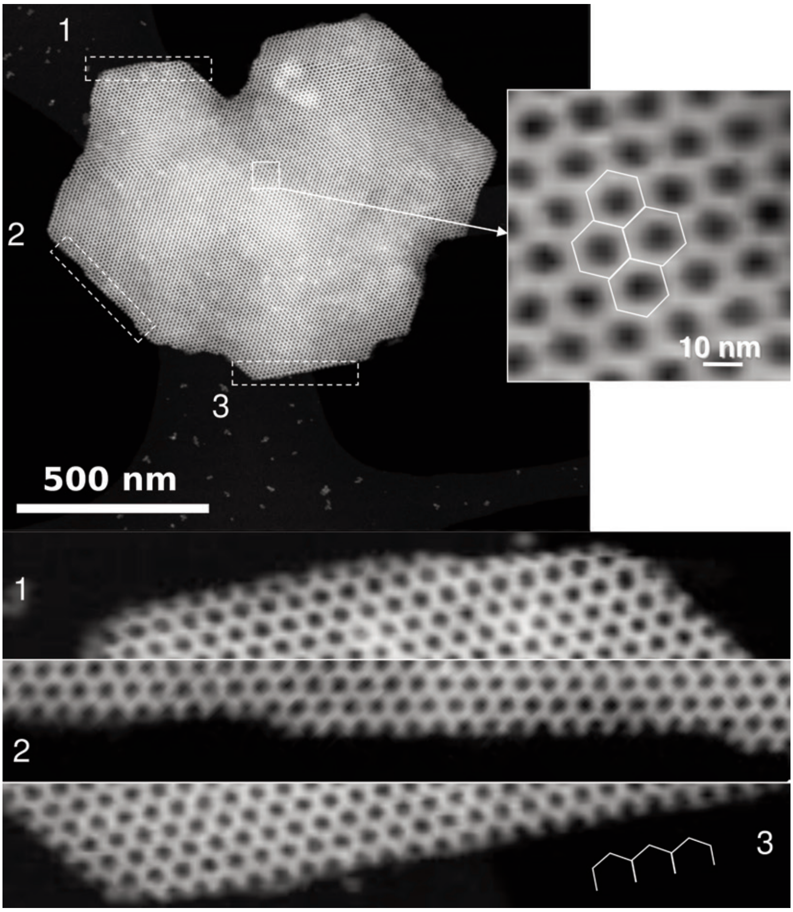
Fortunately, there are now solid catalysts that can be tailored to effect oxidations and other transformations using air or oxygen or hydrogen peroxide as the key reagents. I have myself been involved in fashioning such catalysts. One way of designing powerful new catalysts that effect environmentally responsible and economically important chemical transformations is to learn how to prepare open-structure solids, like the one depicted in Figure 3. Such solids possess enormous internal surface areas; typically a gram of such a mesoporous solid has an area in excess of that of a football field. And by adroit chemical manipulation, one may "place" designed catalytically active centers spaced sufficiently far apart so as to ensure that each active site works independently, like those in an enzyme. A wide and versatile range of chemical synthesis and other processes may now be effected using such open-structure solids.[8] Given that hundreds of thousands of such solids should, in principle, exist, and that only a few hundred or so have been prepared to date, it is safe to predict that many powerful new catalysts will in the future be tailored to effect reactions that yield no by-products and do not generate carbon dioxide.[9]

While it will be perfectly possible to find catalysts that can convert renewable feedstocks (like plants and the oils that they create, such as soyabean, corn, and jatropha) or microalgae (which are much better feedstocks than most living systems) into useful products in a sustainable manner, there are many other pressing scientific problems that must be borne in mind. The most vital of these is the creation of energy to satisfy the rising living standards of a growing world population. The combustion of biomass or the harnessing of wind, hydro, or nuclear forces cannot (even in concert) meet the pressing need for more energy to power our world. The only possible source that is essentially limitless is solar energy; and here new photocatalysts need to be designed and built. This topic has been eloquently discussed by Daniel Nocera in a previous issue of *Dædalus*.[10] He showed, for example, that the energy needs of humans by around 2050 could not possibly be met by nuclear reactors. At present, human activity on Earth requires thirteen terawatts (or thirteen trillion watts) per year to sustain itself. A minimum of ten thousand nuclear fission reactors would have to be built by 2050 to meet the predicted energy needs of the human population, as pointed out by Nathan Lewis.[11] This would require us to build a new nuclear power plant somewhere in the world every other day for the next half-century. At present, the only realistic way in which humankind's energy demands can be met is through harnessing solar radiation, and that will almost certainly necessitate the discovery of new photocatalysts or variants of the nanotechnologic cells involving semiconducting liquid junctions of the kind now under development in various research laboratories.[12]

One may safely predict that, for survival reasons alone, chemists of the future will certainly focus on this pressing problem.[13] But by its very nature, scientific research will see many unexpected discoveries that could well transform future life, so the solutions may not come in the forms we expect (one recalls here the prognostications of the editors of *Scientific American* in 1920). Here, it is prudent to focus on the kind of developments that have come our way in the chemical sciences: examples not of failed predictions but of unexpected successes.

My first example of an unexpected, unpredictable advance was the discovery of green fluorescent protein (which earned Roger Tsien, Martin Chalfie, and Osamu Shimomura the Nobel Prize in Chemistry in 2008). This major advance in biochemistry, which only a few generations ago was inconceivable, owes an enormous debt to Shimomura's lifelong pertinacity and devotion in collecting and studying, out of

Figure 3
An Open-Structure Solid



Various electron tomographical images of mesoporous silica. The nanopores in this solid catalyst show up as dark circles. Their diameter is close to 10 nm (ten billionths of a meter). Source: These micrographs were recorded by my colleagues on samples provided by V. Alfredsson of the University of Lund, using the techniques described in R. K. Leary, Paul A. Midgley, and John Meurig Thomas, "Recent Advances in the Application of Electron Tomography to Materials," *Accounts of Chemical Research* 45 (10) (2012): 1782.

sheer curiosity, 850,000 specimens of jellyfish.

My second example of an unexpected advance has its origins in the borderland between theoretical physics and quantum chemistry. In the 1920s, the young Paul Dirac undertook to study quantum mechanics, stimulated by the work of physicists Werner Heisenberg and Max Born in Germany, and motivated by a desire to incorporate relativistic features into the Schrödinger equation. Dirac's mathematical formulations led him to propose in 1927 the existence of the positron: the first-ever suggestion that antimatter was a reality. It took another four years before the positron's existence was incontrovertibly established through experimental proof by Carl Andersen at the California Institute of Technology. For many decades thereafter, the positron was regarded as a novelty with little prospect of ever being harnessed for practical purposes. Now, however, almost every major hospital in the developed world uses positrons in the noninvasive medical technique of positron emission tomography. Its many uses include charting cerebral activity and identifying stages in the growth of tumors.

My third example of an unexpected scientific leap forward relates to research activities conducted by Allen Hill and his group at the University of Oxford's Inorganic Chemistry Laboratory in the mid-1970s. At that time, Hill began to wonder whether a metalloenzyme could readily exchange electrons with an electrode. In 1982, following up his curiosity-driven question, he invented a simple sensor for glucose in blood, a breakthrough that has subsequently led to the worldwide use of billions of sensor strips by patients suffering from type 2 diabetes (a disease afflicting over 290 million people as of 2010).[14] The measurement of blood glucose levels is required of all patients to whom insulin is prescribed for this disease.

Irrespective of its practical applicability, chemistry as a discipline has a validity of its own. I first realized this fact at the age of eighteen, when I read the first few pages of Sir Cyril Hinshelwood's textbook, *The Kinetics of Chemical Change*, as part of my undergraduate course. His opening paragraph grips me now as it did then:

> That everything changes is an inescapable fact which from time immemorial has moved poets, exercised metaphysicians, and excited the curiosity of natural philosophers. Slow chemical transformations, pursuing their hidden ways, are responsible for corrosion and decay, for development, growth and life. And their inner mechanisms are mysteries into which it is fascinating to inquire.

When Hinshelwood was president of the London Chemical Society more than sixty years ago, he elaborated on the merit of his subject in ways that still resonate today. Not only is chemistry a mental discipline, it is an adventure and an aesthetic experience. Its followers seek to know the hidden causes that underlie the transformations of our changing world; to learn the essence of the rose's color, the lilac's fragrance, and the oak's tenacity; and to understand the secret paths by which the sunlight and air create these wonders. To this knowledge they attach an absolute value: that of truth and beauty. In its pursuit, numerous fascinating, unexpected, and often extraordinary discoveries are made.

As I have amplified elsewhere,[15] scientific researchers know that discoveries cannot be planned: they pop up, like Puck, in unexpected corners.

ENDNOTES

Author's Note: I am grateful for stimulating conversations with J. D. Dunitz, P. P. Edwards, H. A. O. Hill, R. Ritz, P. A. Midgley, and R. K. Leary.

[1] John Kendrew repeated this statement in his book *Thread of Life: An Introduction to Molecular Biology* (London: Bell and Hyman, 1966).

Kendrew's own success owed much to a concatenation of fortunate circumstances and development of new instruments. First, he declined Sir Lawrence Bragg's offer to join him when Bragg became Director of the Royal Institution (RI) in London in 1953. He preferred to stay in Cambridge, but he agreed to visit the RI regularly as Honorary Reader there. This brought him in touch with David Chilton Phillips and Ulrich Wolfgang Arndt, who together had just produced the first-ever so-called linear automatic X-ray diffractometer, a major advance in instrumentation that enormously sped up the collection of X-ray data. But there were two other fortunate occurrences. The first took place in 1951 when Kendrew's gifted research student, Hugh Huxley, was carrying out a series of laborious X-ray–related computations by hand in his Cambridge lodgings, which he shared with an Australian research student named J. M. Bennett. The latter quickly pointed out to Huxley and Kendrew that the unique Electronic Delay Storage Automatic Calculator (EDSAC), designed a few years earlier by Maurice Wilkes in Cambridge, could cope very easily with such calculations. Kendrew and Bennett soon wrote a definitive article ("The Computation of Fourier Syntheses with a Digital Electronic Calculating Machine") that constituted a turning point in the processing of X-ray crystallographic data.

The second fortunate circumstance was that Kendrew was part-time Deputy Chief Scientific Advisor to the Ministry of Defence in London – a position that enabled him to become acquainted with the most powerful computer in Britain, which he used to accelerate the interpretation of his raw data. With the ultra-rapid data collection of the linear X-ray diffractometer and an equally rapid means of processing data, Kendrew and colleagues were able to describe the first-ever three-dimensional model of a protein, myoglobin (the primary oxygen-carrying pigment of muscle tissue), in a 1958 issue of *Nature*. By now, the technique pioneered by Kendrew and Max Perutz (who solved the structure of hemoglobin a few years later) is used extensively worldwide. Some one hundred thousand three-dimensional structures are in the Protein Data Bank, which was created in the 1970s in the Brookhaven National Laboratory and then transferred to Rutgers University. It now contains information about the structures of tens of thousands of proteins and nucleic acids, and it is updated weekly.

[2] There was nothing prophetic in this prediction. The German workers Fischer and Tropsch had already shown in the early 1920s that hydrocarbons for fuels could be prepared, by use of appropriate catalysts, from the products of partially oxidized coal.

[3] In 1984 – 1985, I served as a member of the U.K. Government's Cabinet Office on a panel set up to investigate promising areas of scientific research. Of the many predictions we made, only two became a reality: magnetic resonance imaging (MRI) and the extensive use, in biology and medicine, of confocal light microscopy.

[4] A patent for a jet engine was taken out in the early 1930s by the young British flying officer Frank Whittle.

[5] The Hindenberg disaster in 1927 put an end to the popularity of long-distance travel by dirigible.

[6] J. A. Ballantine, J. H. Purnell, and J. M. Thomas, "Sheet Silicates: Broad Spectrum Catalysts for Organic Synthesis," *Journal of Molecular Catalysis* 27 (1) (1984): 157.

[7] See John Meurig Thomas, "Solid Acid Catalysts," *Scientific American* 266 (4) (1992): 112.

[8] John Meurig Thomas, *Design and Applications of Single-Site Heterogeneous Catalysts: Contributions to Green Chemistry, Clean Technology and Sustainability* (London: Imperial College Press, 2012).

[9] John Meurig Thomas and Jacek Klinowski, "Systematic Enumeration of Microporous Solids: Towards Designer Catalysts," *Angewandte Chemie International Edition* 46 (38) (2007): 7160.

[10] Daniel G. Nocera, "On the Future of Global Energy," *Dædalus* 135 (4) (2006): 112. See also Daniel G. Nocera, "Can We Progress from Solipsistic Science to Frugal Innovation?" *Dædalus* 141 (3) (2012): 45.

[11] Nathan S. Lewis, "Powering the Planet," *MRS Bulletin* 32 (2007): 808.

[12] Ibid.

[13] Predictions in general have exercised the thoughts of numerous historians of science as well as those of business-school economists. See, for example, Stephen G. Brush, "Prediction and Theory Evaluation: The Case of Light Bending," *Science* 246 (4934) (1989): 1124; Clayton M. Christensen, *The Innovators' Dilemma: When New Technologies Cause Great Firms to Fail* (Boston: Harvard Business School Press, 1997); and Norman R. Augustine, "They Never Saw It Coming," *Science* 339 (6118) (2013): 373.

[14] Anthony E. G. Cass, Graham Davis, Graeme D. Francis, H. Allen O. Hill, William J. Aston, I. John Higgins, Elliot V. Plotkin, Lesley D. L. Scott, and Anthony P. F. Turner, "Ferrocene-Mediated Enzyme Electrode for Amperometric Determination of Glucose," *Analytical Chemistry* 56 (4) (1984): 667; and Jane E. Frew and H. Allen O. Hill, "Electrochemical Biosensors," *Analytical Chemistry* 59 (15) (1987): 933A.

[15] John Meurig Thomas, "Intellectual Freedom in Academic Scientific Research Under Threat," *Angewandte Chemie International Edition* 52 (22) (2013): 5654.

The Bright Future of Fabulous Materials Based on Carbon

Fred Wudl

Abstract: Our current civilization belongs to the organic materials age. Organic materials science pervades nearly all aspects of our daily life. This essay sketches the evolution of materials science up to the present day. Plastics as textiles and structural materials dominate human civilization. The element carbon is at the core of this development because of its diverse interconnections with itself and other elements of the periodic table. While silicon will not be supplanted from its role in electronics, carbon will provide the most versatile electronics applications, through inexpensive, flexible electronic devices.

> Mr. McGuire: Ben, just one word.
> Ben: Yes, sir?
> Mr. McGuire: Are you listening?
> Ben: Yes, I am.
> Mr. McGuire: Plastics!
> Ben: How do you mean?
> Mr. McGuire: There is a great future in plastics, think about it.
>
> – *The Graduate* (1967)

Mr. McGuire was right in his assessment. Plastics, or better yet, organic materials, contribute significantly to making our everyday lives exciting and productive. It is hard to imagine a world without plastics. Over the last several decades, organic materials have given us polyesters, polystyrene, formica, lightweight containers, fiberglass boats, airplanes that consume less fuel (thanks to lighter structural materials), fibers that are stronger, nylon, and much more recently, flat screen displays with brilliant, vibrant colors. And organic materials are today leading us into a much less energy-intensive future, thanks to organic solar cells ("plastic solar cells") and organic transistors.

How have we come so far so quickly: from unimaginable to reality in as little as sixty years? Let's

FRED WUDL, a Fellow of the American Academy since 2001, is Research Professor of Chemistry and Materials at the University of California, Santa Barbara. His current research interests include the optical and electro-optical properties of processable conjugated polymers, the organic chemistry of fullerenes, and the design and preparation of self-mending polymers.

doi:10.1162/DAED_a_00303

backtrack to examine this history a bit more closely.

Materials science might well be labeled the "central science," a moniker given to the discipline of chemistry in the recent past.[1] Materials predate chemistry by millennia, and human civilization is inextricably connected with materials. In fact, the stages of early human civilization have been characterized by the materials that humans used as civilization evolved. Thus the progress in prehistory follows the sequence stone age (Neolithic), followed by bronze age, and then iron age. Thereafter (c. 5000 BCE), civilization became rather complex; but if we were to jump forward to the present day, we would find that *every aspect of current civilization is touched by electronics*, and specifically, electronics based on highly processed silicon. This brings us to *electronic materials* that have been dominated by semiconductor elements, in particular silicon, germanium, and compound semiconductors such as gallium arsenide, indium gallium arsenide, and so on. In parallel with the evolution of solid-state electronics, organic materials (particularly plastics) have fully permeated civilization, hence a proper name for the current civilization, in terms of materials, would be the *silicon age*, or the *organic materials age*. The much more modern aspect of materials science, namely, *organic electronic materials*, is in its nascency; its origin can be traced to the second half of the twentieth century, specifically to 1973[2] and 1987.[3]

Inherent in the definition of organic materials is the classic definition of organic chemistry as the chemistry of carbon compounds. The foundation of organic materials is based on the unique way in which carbon atoms interact with one another and with other elements of the periodic table. The key difference, at the atomic level, between carbon and all the elements below it in its column of the periodic table (that is, silicon, germanium, tin, and lead) is its ability to form *stable double bonds* (π bonds), *particularly with itself*. The π bonds consisting of two pairs of shared electrons are the lowest common denominator in organic electronic materials. Another fundamental difference between organic solids and inorganic solids (again at the atomic level) is that the former are *molecular solids* while the latter are *extended solids*. Extended solids are materials in which the entire bulk has all atoms bonded to each other: for example, a chunk of silicon, a diamond, a gold nugget, iron, silicon oxide (quartz, glass), or iron oxide (magnetite). On the other hand, molecular solids are composed of discrete molecules that are not bonded to each other. The molecules themselves can consist of a diverse number of atoms, ranging from two (iodine) to millions (ultra high molecular weight poly[ethylene], VECTRA, DNA). The molecules pack into a solid in which intermolecular forces hold the bulk material together. Intermolecular attractive forces are many orders of magnitude weaker than the interatomic bonds of solids. Thus, molecular solids in general are soluble in most solvents and have relatively low melting points, whereas *all* extended solids are insoluble in all solvents, and with the exceptions of mercury, cesium, gallium, rubidium, and potassium, as well as their alloys, extended solids have rather high melting points.

The obvious advantage of organic materials is that they are *much easier to process (that is, convert to useful items) than are traditional inorganic materials*. Thus, organic materials are considerably less energy intensive in the procedures employed to convert them to useful objects. Organic materials have had profound effects on society via their roles as *textiles*, *structural materials*, and very recently, *electronic materials*. The latter will be the main subject of this essay.

Organic materials in textiles had their start with natural flax fibers going back to the Neolithic age, roughly thirty thousand years ago;[4] this was followed by cotton[5] at least seven thousand years ago. It was not until the beginning of the twentieth century that synthetic, organic polymer–based ("plastics") fibers became prevalent. They had originated in the mid-nineteenth century with celluloid, the first film and bulk plastic-parts-forming material. Celluloid had its origin with nitrocellulose, the explosive guncotton. In 1855, in Birmingham, England, Alexander Parkes was the first to convert nitrocellulose into a plastic by mixing a nitrocellulose solution, known as collodion, with camphor and then evaporating the solvent. He named his invention Parkesine, and the material is thought to mark the start of the plastics industry.[6]

By the 1870s, it became clear that certain objects made from elephant tusk ivory, such as billiard balls and piano keys, were becoming very expensive. Enter John Wesley Hyatt, an entrepreneur and inventor who expanded on Parkes's invention and labeled his material celluloid. A myriad of articles were made of celluloid between the second half of the nineteenth century and the early twentieth century, including cinematographic film. Not surprisingly, the film was very unstable and almost "spontaneously" combustible, resulting in tragic movie theater fires. These semi-synthetic plastics had a limited lifetime, either suffering discoloration or breaking down. In 1907, Leo Baekeland invented a synthetic resin made out of inexpensive synthetic materials (phenol and formaldehyde) and called it Bakelite. Bakelite was stronger than celluloid, was very stable, and was not nearly as combustible. Soon a wide range of everyday objects were made of Bakelite: fountain pens, telephone housings, knife handles, art deco objects, and phonograph records, among many others.

In 1924, another semi-synthetic material was obtained from wood cellulose processed into rayon fibers,[7] yet very little was known about these materials at either the atomic or molecular level. Many chemists were convinced that plastics consisted of another form of matter, one that did not involve traditional chemical bonds. It was not until the fundamental contributions of Herman Staudinger at the turn of the twentieth century[8] and Paul Flory[9] in the mid-twentieth century that we were able to explain all the properties of plastics as resulting from the properties of giant molecules (macromolecules or polymers). Both Staudinger and Flory received Nobel Prizes (in 1953 and 1974, respectively) for their contributions to polymer chemistry. Nylon and all its successors (for example, polyester, acrylic, saran, and spandex) were invented based on our understanding of the chemistry of macromolecules at a fundamental level.

The terms *polymer* and *macromolecule* have become synonymous, while plastic usually refers to a material made of synthetic polymers. As the name implies, a polymer consists of many (Greek, *poly*) identical units or parts (Greek, *mer*) joined together. The simplest plastics are those in which a long molecular chain is formed of repeat units (chain links) that are all identical (homopolymers), as is the case with polyethylene, polypropylene, and polystyrene. One can imagine that because carbon is so versatile in its bonding ability, the potential for different homopolymers is almost infinite. Now imagine combining two different monomers to form a molecular chain. They could be joined in at least three different types of arrangements: they could alternate, occur at random, or occur in blocks. Each of these arrangements can be achieved, and each results in *designable* properties. Chemists and materials scientists, thanks to carbon's unique chemistry, have at their disposal an almost infinite number of organic plastic materials that lend themselves to the production of everything from rubber bands to airplanes.

Although the natural polymers making up linen, cotton, and wool are still major textile fibers, the highest strength fibers, as well as those with special applications, are all synthetic. For the same weight, a Kevlar fiber is stronger than steel. How is it possible to have a material that, in fiber form, is stronger than steel and yet is a molecular solid? There are two reasons: 1) the macromolecules are very long, resulting in their forming regions of entanglements or "mechanical bonds"; and 2) the extensive regions that are not entangled are highly ordered and have enhanced intermolecular attractive forces. These intermolecular attractions are hydrogen bonding, dipole-dipole, π-π stacking, and Van der Waals attractive forces. They were known to chemists for decades but were not exploited until the second half of the twentieth century, at which point organic chemists were able to design molecules exploiting these weaker attractive forces. These "designer molecules" would then order themselves into a predetermined structure by "self assembly."

Kevlar was designed to maximize all these intermolecular forces, starting with the monomer, the method to link the monomers to each other (polymerization), and finally the processing into fibers. The chain entanglements are the main reason why polymeric materials are plastic (flexible, malleable, and ductile), and hence convertible into fibers, films, devices, and machine parts. The latter are in the realm of structural or engineering materials.

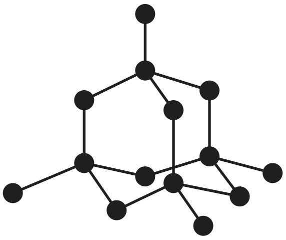
> Our everyday objects are lighter, more durable, cheaper to manufacture, sleeker and more sanitary than their predecessors. The world before plastics was that of butcher paper, tin soldiers, wooden crates, broken glass and rusting metal – less convenient and less abundant in consumer goods than today.... Engineering polymers – developed for the purpose of replacing metals – made up the last major class of materials to spring from the Golden Age of Plastics.
>
> – Alexander H. Tullo, *Chemical & Engineering News*, September 9, 2013

The mechanically stronger and, in general, more brittle plastics such as Bakelite or Formica have their long molecules connected to each other, with interchain bonds or crosslinks. In essence they are a form of extended solids. Hence, they are insoluble and infusible. The only way to process them is by forming the part from its monomer components in a mold through a method called *thermosetting*. Because polymeric solids are so easily manipulated, by further chemical transformation, by dissolution, or by melting, they can also be converted to very strong and tough materials, particularly by forming composites with strengthening ingredients such as glass fibers, carbon fibers, or steel mesh. Our most advanced airplanes (the F-117 stealth fighter, the B-2 bomber, the Boeing 787 Dreamliner) now feature extensive use of plastic-carbon fiber composites.

The giant molecules used in structural materials, while often containing π bonds, do not take advantage of the π bonds' exquisite electronic properties. Thanks to the fundamental research of organic chemists, mostly in the twentieth century, π bonds can be tailored to have a useful electronic property and this ability is the essence of organic electronic materials. An isolated double bond between two carbon atoms is relatively uninteresting from an organic electronics perspective because the energy required to manipulate these π electrons is too high for electronic devices. Figure 1 shows that as π bonds are connected, the energy required to excite their electrons becomes smaller. Thus the wavelength of light required to excite electrons from the relatively high-energy ultraviolet region of the spectrum for ethylene (one π bond) to the lower energy visible region in carotene (eleven π bonds). By increasing the number of alternating π bonds in a molecule and including other elements that interact with the electrons of π bonds (oxygen, sulfur, nitrogen), one can extend the absorption of light to wavelengths of light corresponding to relatively low energies (to the near infrared region of the spectrum).

Ethylene and all other π-bonded carbon atoms are planar: that is, all four hydrogen atoms of ethylene are in the same plane; deforming the plane results in a highly "strained" bond. The most extremely interconnected π-bonded carbon structure known is graphene, a small section of which is shown in Figure 2.

In nature, carbon occurs in two major "allotropes": diamond and graphite. The former, as mentioned above, is an extended three-dimensional solid. The structure of a very tiny piece of diamond is shown below.



In this structure, each carbon atom is joined to four other carbon atoms by single bonds. On the other hand, graphite is composed of a very large number of planar graphene sheets stacked on top of each other, with perfect registry from graphene layer to layer, held together by weak Van der Waals forces. Each sheet contains many bonds. Because these inter-graphene forces are so weak, in writing with a graphite pencil, one scrapes off layers of graphene onto the paper. In 2010, the Nobel Prize in Physics was awarded to A. K. Geim and K. S. Novoselov[10] for their discovery of the very unusual properties of graphene.[11] Graphene can in fact be considered the conceptual germ of two other new forms of carbon, buckminsterfullerene[12] and carbon nanotubes, illustrated in Figure 3.[13]

For the discovery of fullerenes, R. F. Curl, H. W. Kroto, and R. E. Smalley received the Nobel Prize in Chemistry in 1996. Fullerenes, especially carbon nanotubes, can be envisioned as arising from the curving of a graphene sheet onto itself. The electronic properties of buckminsterfullerene and carbon nanotubes are directly related to the strain of their bent π bonds. In buckminsterfullerene, the strain is manifested by an increase in electronegativity, meaning that each π bond becomes mildly electron-attracting, making buckminsterfullerene an electron acceptor (EA) molecule. With the advent of buckminsterfullerene, chemists were presented for the first time with a set of π bonds arranged on a spherical surface. Buckminsterfullerene is also the first carbon allotrope in the form of a molecular solid that is soluble in several solvents and sublimable under vacuum at readily accessible temperatures. Buckminsterfullerene molecules tend to aggregate, a property that is auspicious for the development of organic solar cells, as discussed below.

Buckminsterfullerene and carbon nanotubes are *at the heart of organic nanoscience and nanotechnology,* a relatively new branch of chemistry, physics, and materials science.[14] Nanotechnology and nanoscience deal with properties of matter in the realm of 1 – 100 nm. In this scale, properties may be dominated by quantum mechanical effects. Two typical examples (of many)[15] of current nanotechnology applications in materials science are bandages impregnated with silver nanoparticles, to take advantage of their remarkable antibiotic properties, and nano-sized titanium dioxide as well as zinc oxide for particularly effective sun screen ointments.[16] A truly up-to-date

Figure 1
Carotene, with Its Eleven Linked (Conjugated) π Bonds

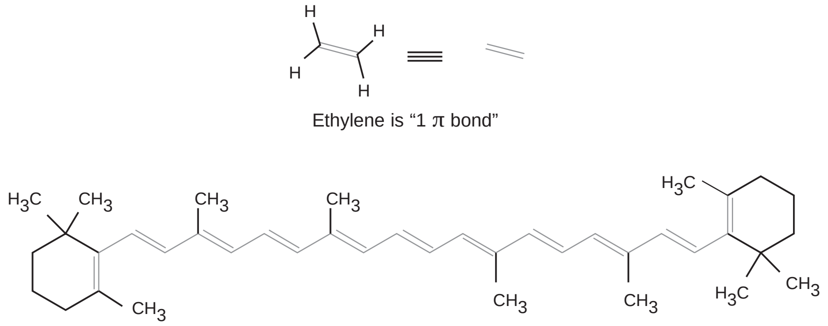
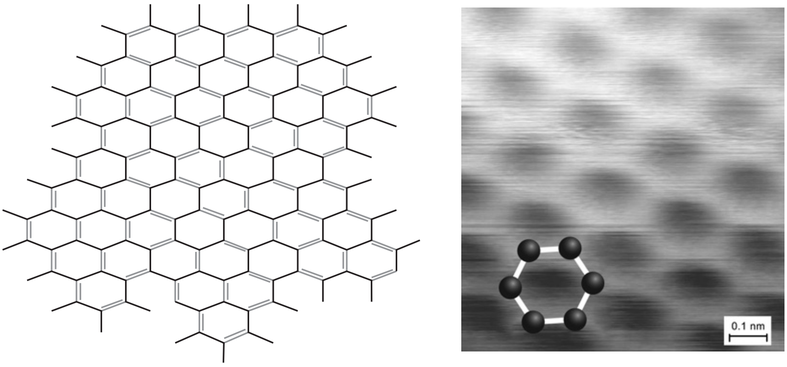


Figure 2
The π Bonds in Graphene (left); an Atomic Force Micrograph of Graphene (right)



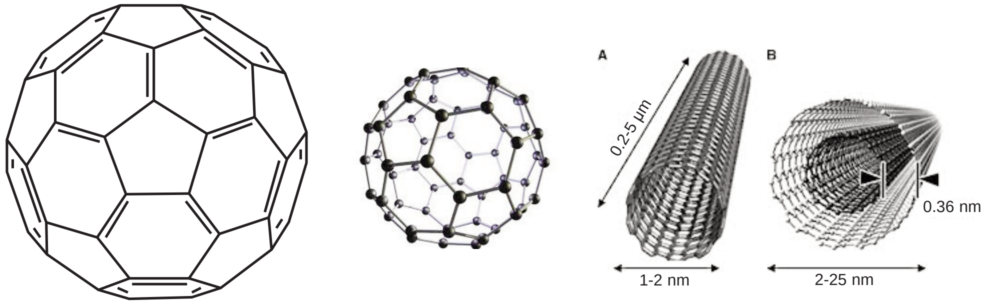
One nanometer (nm) is approximately one ten-billionth of an inch. Source: E. Stolyarova, K. T. Rim, S. Ryu, J. Maultzsch, P. Kim, L. E. Brus, T. E. Heinz, M. S. Hybertsen, and G. W. Flynn, “High-Resolution Scanning Tunneling Microscopy Imaging of Mesoscopic Graphene Sheets on an Insulating Surface,” *Proceedings of the National Academy of Sciences* 104 (2007): 9209–9212.

application of carbon nanotubes is in computing, where nanotubes are beginning to take the place of silicon in a computer’s integrated circuit.[17]

To make significant advances in organic electronics, scientists need to be truly interdisciplinary. Equal participation of chemists, physicists, and engineers is required. The chemists design and synthesize the organic materials, the physicists provide the theory and experimental procedures for transporting electrons through organic molecular solids, and the engineers provide the design and fabrication of devices for the consumer.

The close collaboration of chemists and physicists helped establish that the most effective way to generate and delocalize electrons in organic solids is to use two types of π bond–containing molecules: electron acceptors (EA) and electron donors (ED).[18] We already saw that buckminsterfullerene is an electron acceptor. Electron donor molecules have π electrons that are relatively loose and easily given up, leaving behind a positive charge or “hole.” The solid resulting from transferring an

Figure 3
Structure of Buckminsterfullerene with Sixty Carbon Atoms (C_{60}) Containing Thirty π Bonds (left); Artist's rendering of a Single and Multiwall (Double) Carbon Nanotube (right)

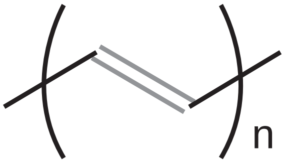


Source: http://en.wikipedia.org/wiki/File:Buckminsterfullerene-2D-skeletal.png; A. Hirsch, "Funktionalisierung von einwandigen Kohlenstoffnanoröhrern," *Angewandte Chemie* 114 (2002): 1933–1939; R. M. Reilly, "Carbon Nanotubes: Potential Benefits and Risks of Nanotechnology in Nuclear Medicine," *Journal of Nuclear Medicine* 48 (7) (2007): 1039–1042; and S. Iijima, "Carbon Nanotubes: Past, Present, and Future," *Physica B* (323) (2002): 1–5.

electron from an ED to the EA is known as "a charge transfer complex." It was further established that both ED and EA molecules had to form infinite pancake-like stacks in the solid state. Electrons and holes travel along these stacks much more readily than they do between stacks. With this fundamental knowledge, scientists were able to invent organic solids that conduct electricity just as well as many metals do (but still not as well as copper). Another property that these organic solids shared with metals was that their conductivity *increased* with decreasing temperature and as a result, these solids were known as *organic metals*.

Up to the time of this momentous discovery, organic materials were useful in electrical engineering and electronics because they were excellent *insulators*, not conductors. Prior to this breakthrough point in the history of organic electronics,[19] the measurement of conductivity as a function of temperature of organic solids afforded only very low conductivity that *decreased* with decreasing temperature, a property that defines semiconductors in general. Further research on organic metals ultimately led to the discovery of organic superconductors. These are materials that exhibit infinite conductivity (zero resistance) below a particular transition temperature.[20] This was a truly amazing development because up to that point it was believed by all experts in the field that in order to observe superconductivity, one needed an extended solid.

From a fundamental point of view, achieving metallic and superconducting properties with organic materials was clearly a remarkable accomplishment. But from a materials engineering viewpoint this was not the case, because organic metals were tiny, brittle crystals that could not be processed into useful devices. What was required was a polymer or plastic that would exhibit high conductivity when converted to a charge transfer complex. Not until 1977 did this momentous discovery take place, when collaborating scientists in Japan and the United States identified an electrically conducting polymer.[21] This finding led to the 2000 Nobel Prize in Chemistry being awarded to A. J. Heeger, A. MacDiarmid, and H. Shirakawa. From an organic chemist's perspective, this was a deceptively simple polymer, a polyacetylene. It was simply a solid made up of molecules that consisted of long chains of conjugated π bonds (where *n* is a difficult-to-determine large number in the hundreds to thousands).



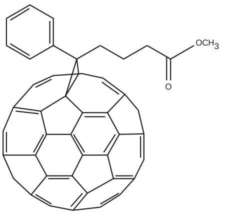
A sample of polyacetylene looks like a piece of aluminum foil, but it tarnishes quickly when exposed to the atmosphere. This material was chiefly of academic interest, but it provided a trove of fundamental information, much like fruit flies do for the field of genetics. In quick succession, the organic electronics community developed much more stable polymers that could be useful as materials. The paradigm of the day was to achieve ever higher conductivities, even conductivities that might be achievable without having to resort to charge transfer complex formation. As a result, the chemistry and physics community concentrated on achieving metal-like properties until Richard Friend's group (in 1990) discovered that one of the most stable polymers could be processed into a light-emitting diode.[22] This was a semiconductor property, not a metal property. A light-emitting diode is a particular semiconductor device that emits light when a voltage is applied across it.

The discovery caused an immediate paradigm shift by the research community, from the search for metal properties to the search for semiconductor properties. In fact, judging from the number of publications, research on organic metals appears to have ceased in the United States, although it is still pursued in Japan and Europe. While the discovery of the polymeric light emitting diode (PLED) was clearly an important step in the development of organic electronics, the first organic electronics device that was reported came in 1987;[23] it was also a light-emitting device, but was not based on polymers. Rather, it was a small-molecule organic light-emitting diode (OLED). Both kinds of devices are very bright and colorful, but so far, PLEDs are still in the development stage while OLEDs have advanced into the commercial sphere, seen everywhere from flat screen display devices to smartphones.

Because electrons and holes carry opposite charges, they attract each other and can actually combine, resulting in loss of charge carriers. The origin of the light emitted by OLEDs and PLEDs is based on the fact that when electrons recombine with holes, they raise molecules to a high-energy excited state. When the excited molecules return to a ground state they emit light or heat.

Again, π bonds have a greater tendency to emit light than heat. In OLEDs and PLEDs, the positive and negative electrodes of the diode create the holes and electrons, respectively. The excited state of an electron-hole pair is called an *exciton*. An exciton can also be created by absorption of light of the wavelength corresponding to the energy required to produce an electron-hole pair. So, if one were able to separate the holes from the electrons before they had a chance to recombine, then one would create an electric field or voltage by the absorption of light, and one would have a "photovoltaic device" or "solar cell" if the light absorbed were that corresponding to the solar spectrum. Many attempts to produce organic photovoltaic cells were made throughout the period from the 1960s to 1980s by combining EDs and EAs and irradiating them. Unfortunately, though the devices produced electricity, they had extremely low efficiencies, mostly because the electron-hole recombination rates were too high. Another way to view this observation is that the acceptors were "too willing" to give the electron back to the holes.

As it turned out, the negatively charged fullerenes, resulting from accepting an electron, were less strained than the slightly smaller-diameter neutral fullerenes. This property resulted in a one thousand to ten thousand times slower recombination[24] with the hole of an exciton generated in an ED polymer than was observed previously with any other EA. Even though the first organic solar cells – "plastic solar cells" – based on fullerene, discovered in 1992, had much higher efficiencies (0.04 percent)[25] in the conversion of solar energy to electrical energy (power conversion efficiency [PCE]) than any previous organic device, it was still far too low to be of more than academic interest. Another important aspect of fullerenes for this application was that the very tight aggregates fullerenes tend to form allowed the negative charge to be carried quickly toward the negative electrode of the cell. With concentrated research efforts in the United States, Europe, Japan, and China, the PCE increased rapidly to the current record of 11 percent,[26] a 275-fold increase from the original efficiency. While this number may seem small, keep in mind that solar cells based on amorphous silicon exhibit a PCE of 10 to 11.9 percent. The most efficient solar cell reported to date is 38.8 percent efficient, and commercial cells are only 11 to 19 percent efficient.[27] The most efficient non-organic experimental cell is reported to be 44.4 percent efficient, but this involves the use of a solar concentrator, a device, like a magnifying lens, that concentrates light; without a concentrator, the maximum PCE is 32.6 percent.[28] There are now several small companies manufacturing plastic solar cells that are adequately efficient for relatively small applications, such as on the roof of a bus shelter for the minimal electricity needs there (lights, for example), or on the outer flap of a woman's purse for the powering of a cell phone. None of these devices use fullerene C_{60} as the acceptor, but rather a derivative called PCBM (shown below), an acronym for [6,6]-phenyl-C_{61}-butyric acid methyl ester.



While fullerene is relatively insoluble, its electronic structure makes it slightly too strong of an electron acceptor, whereas PCBM is more soluble and is a slightly weaker electron acceptor, better matching the electron-donating properties of most polymeric EDs. The development of PCBM provides yet another example of how many times science progresses by serendipity. This fullerene derivative was originally prepared as part of a program to make a water-soluble agent to inhibit the active site of HIV protease, an enzyme used by the AIDS virus in its replication process. The active site of the protease is a cavity of approximately 1 nm in diameter, corresponding to the diameter of fullerene. To be able to examine the biological properties of any molecule, the molecule needs to have some water solubility. Fullerene is insoluble in water, and PCBM was prepared as a water-soluble fullerene derivative. Another reason that fullerene PCBM was prepared is simply because the chemistry of fullerene was being explored in my research group as well as in several other groups around the world; and one of the better known methods to transform fullerene was by the reaction that led to PCBM. At the same time, the "materials" properties of fullerene, particularly the optoelectronic properties for organic electronics, were being examined by Dr. Saiciftci in Alan Heeger's group at the University of California, Santa Barbara. He wanted a soluble fullerene derivative because he had just determined that fullerene would accept an electron from a photoexcited ED polymer. Changing from fullerene as an electron acceptor to PCBM dramatically enhanced the PCE to 2.9 percent.[29] From 1995 to 2013, the PCE improved to 11 percent. This significant enhancement was a consequence of a change in the ED polymer architecture from homopolymers to alternating co-polymers. This structural modification allowed for more versatile design of electronic character of the π backbone of the polymer.

In order to fabricate a conventional silicon-based solar cell one must first process the silicon. Because the melting point of silicon is 1414 degrees C (2577 degrees F), and because in order to fabricate silicon-based devices one must crystallize silicon from the melt, the manufacture of solar cells is a very energy-intensive process. On the other hand, plastic solar cells are based on molecular solids. They can be processed from solution, called "inks," with very simple devices such as a dot matrix printer or a roll-to-roll printer. Thanks to their light weight, flexibility, ease of manufacture, and low cost, plastic solar cells can be expected to have a profound effect on the world's non-fossil fuel, non-nuclear electricity generating capacity. A similar bright future for the reduction of energy-consumption can be foreseen for OLED display devices, since these devices do not need strong backlighting. To observe a non-luminescent device, such as a liquid crystal display (LCD), one needs a source of illumination; for flat screens this is best accomplished from the back, hence "backlighting."

As stated above, when π bonds are not in charge transfer complexes, they behave as semiconductors, the backbone of modern electronics. Diodes and transistors are the simplest units of semiconductor electronic devices. We already saw an organic diode application in the form of OLEDs. Transistors are solid-state microscopic switches and amplifiers that are the main way to control electron flow in electronic devices, particularly in computing and displaying. Organic electronics is influencing the electronics industry with organic transistors. The easiest way to convert an organic semiconductor to a transistor is to use the tendency of organic materials to readily form thin films. Organic transistors are thin film transistors (TFT). A particularly simple TFT is the field effect transistor (FET); in this case an organic FET would be an OFET.[30] Pentacene, a 22-π-bonded carbon molecule, its derivatives, and some polymeric materials that are much better film-formers easily outperform amorphous silicon and have already yielded practical devices, such as drivers for flexible OLED and liquid crystalline displays.

In conclusion, by exploiting the π bonding capability of carbon, organic materials have been designed to exhibit unmatched advantages: they are lightweight, flexible, and low-cost; and they have low energy fabrication demands. Devices based on these materials can be expected to proliferate rapidly. Cheaper and lighter large-scale structures (dwellings and bridges, for instance); countless electronic devices; cheaper, lighter, and recyclable containers; more efficient modes of transportation; and improved techniques for the harvesting of solar energy: all will be part of this branch of organic chemistry's contribution to sustainable life on Earth.

ENDNOTES

Author's Note: This essay has benefited from the feedback of Linda Wudl and Jerrold Meinwald, to whom I am indebted for their kind help.

[1] Attributed to Theodore E. Brown, who wrote a freshman chemistry text titled *Chemistry, the Central Science*.

[2] J. Ferraris, D. O. Cowan, V. V. Walatka, Jr., and J. H. Perlstein, "Electron Transfer in a New Highly Conducting Donor-Acceptor Complex," *Journal of the American Chemical Society* 95 (1973): 948 – 949; and F. Wudl, D. Wobschall, and E. J. Hufnagel, "Electrical Conductivity by the Bis-1,3-dithiole-Bis-1,3-dithiolium System," *Journal of the American Chemical Society* 94 (1972): 670 – 672.

[3] C. W. Tang and S. A. Van Slyke, "Organic Electroluminescent Diodes," *Applied Physics Letters* 51 (1987): 913 – 915.

[4] Eliso Kvavadze, Ofer Bar-Yosef, Anna Belfer-Cohen, Elisabetta Boaretto, Nino Jakeli, Zinovi Matskevich, and Tengiz Meshveliani, "30,000-Year-Old Wild Flax Fibers," *Science* 325 (5946) (2009): 1359; and Richard Harris, "These Vintage Threads Are 30,000 Years Old," *All Things Considered*, September 10, 2001, NPR.

[5] Christophe Moulherat, Margareta Tengberg, Jerome-F. Haquet, and Benoit Mille, "First Evidence of Cotton at Neolithic Mehrgarh, Pakistan: Analysis of Mineralized Fibers from a Copper Bead," *Journal of Archaeological Science* 29 (12) (2002): 1393 – 1401.

[6] P. C. Painter and M. M. Coleman, "Essentials of Polymer Science and Engineering," DEStech Publications, Pennsylvania, 2009.

[7] Ibid.

[8] Hermann Staudinger, "Über Polymerisation," *Berichte der Deutschen Chemischen Gesellschaft* 53 (6) (1920): 1073 – 1085.

[9] Paul Flory, *Principles of Polymer Chemistry* (Ithaca, N.Y.: Cornell University Press, 1953).

[10] http://www.nobelprize.org/nobel_prizes/physics/laureates/2010/.

[11] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric Field Effect in Atomically Thin Carbon Films," *Science* 306 (5696) (2004): 666 – 669.

[12] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, "C_{60}: Buckminsterfullerene," *Nature* 318 (6042) (1985): 162 – 163, Bibcode:1985Natur.318..162K, doi:10.1038/318162a0.

[13] Sumio Iijima "Helical Microtubules of Graphitic Carbon," *Nature* 354 (6348) (1991): 56 – 58, Bibcode:1991Natur.354...56I, doi:10.1038/354056a0.

[14] Ed Regis, *Nano: The Emerging Science of Nanotechnology* (Boston: Little, Brown and Company, 1995).

[15] Ibid.

[16] M. E. Kurtoglu, T. Longenbach, P. Reddington, and Y. Gogotsi, "Effect of Calcination Temperature and Environment on Photocatalytic and Mechanical Properties of Ultrathin Sol–Gel Titanium Dioxide Films," *Journal of the American Ceramic Society* 94 (4) (2011): 1101 – 1108, doi:10.1111/j.1551-2916.2010.04218.x.

[17] Max M. Shulaker, Gage Hills, Nishant Patil, Hai Wei, Hong-Yu Chen, H.-S. Philip Wong, and Subhasish Mitra, "Carbon Nanotube Computer," *Nature* 501 (26 September 2013): 526 – 530, doi:10.1038/nature12502.

[18] Regis, *Nano: The Emerging Science of Nanotechnology*.

[19] Ferraris et al., "Electron Transfer in a New Highly Conducting Donor-Acceptor Complex," 948 – 949; and Wudl, "Electrical Conductivity by the Bis-1,3-dithiole-Bis-1,3-dithiolium System," 670 – 672.

[20] J. M. Williams, J. R. Ferraro, R. J. Thorn, K. D. Carlson, U. Geiser, H. H. Wang, A. M. Kini, and M.-H. Whangbo, *Organic Superconductors* (Upper Saddle River, N.J.: Prentice Hall, 1992).

[21] H. Shirakawa, E. J. Lewis, A. G. MacDiarmid, C. K. Chiang, and A. J. Heeger, "Synthesis of Electrically Conducting Organic Polymers: Halogen Derivatives of Polyacetylene, $(\mathrm{CH})_x$," *Journal of the Chemical Society, Chemical Communications* (16) (1977): 578.

[22] J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burns, and A. B. Holmes, "Light-Emitting Diodes Based on Conjugated Polymers," *Nature* 347 (1990): 539.

[23] Tang, "Organic Electroluminescent Diodes."

[24] N. S. Sariciftci, L. Smilowitz, A. J. Heeger, and F. Wudl, "Photoinduced Electron Transfer from a Conducting Polymer to Buckminsterfullerene," *Science* 258 (1992): 1474.

[25] N. S. Sariciftci, D. Braun, C. Zhang, V. I. Srdanov, A. J. Heeger, G. Stucky, and F. Wudl, "Semiconducting Polymer-Buckminsterfullerene Heterojunctions: Diodes, Photodiodes, and Photovoltaic Cells," *Applied Physics Letters* 62 (1993): 585.

[26] M. C. Scharber and N. S. Sariciftci, "Efficiency of Bulk-Heterojunction Organic Solar Cells," *Progress in Polymer Science* (2013), doi:10.1016/j.progpolymsci.2013.05.001.

[27] http://www.nrel.gov/ncpv/images/efficiency_chart.jpg.

[28] Ibid.

[29] G. Yu, J. Gao, J. C. Hummelen, F. Wudl, and A. J. Heeger, "Polymer Photovoltaic Cells: Enhanced Efficiencies via a Network of Internal Donor-Acceptor Heterojunctions," *Science* 270 (1995): 1789.

[30] Ioannis Kymissis, *Organic Field Effect Transistors: Theory, Fabrication and Characterization* (New York: Springer, 2009).

The Convergence of Chemistry & Human Biology

Chaitan Khosla

Abstract: Over the past two decades, "chemical biology" has emerged as the term of choice to describe the interface between chemistry and biology. As its name suggests, the field draws upon chemical insights and tools to understand or engineer living things. This essay focuses on the scientific, societal, and pedagogical potential of an emerging frontier for chemical biologists: namely, the study of Homo sapiens. My goal is to highlight the opportunities and challenges presented to chemistry by human biology at a time when it costs less to sequence an individual's genome than it does to buy a car. But how does chemical biology differ from other similar-sounding fields? By first reaching a clear understanding of the scope of chemical biology, we may address more pertinent questions such as: What is the promise of the emerging interface between chemistry and human biology? Why is it important to nurture the relationship between these fields? And what are the attributes of individuals and environments that are well poised to contribute significantly to this interface?

CHAITAN KHOSLA, a Fellow of the American Academy since 2007, is Professor of Chemistry, Chemical Engineering, and (by courtesy) Biochemistry at Stanford University. He also directs Stanford ChEM-H, an institute that brings together chemists, engineers, biologists, and clinicians to understand life at a chemical level and apply that knowledge to improving human health. His work on antibiotic biosynthesis and the molecular basis for celiac disease has been published in *Science, Journal of the American Chemical Society*, and *Proceedings of the National Academy of Sciences*, among many others.

To someone only vaguely familiar with disciplines such as biochemistry, structural biology, molecular biology, and medicinal chemistry, the introduction of yet another related name may seem unnecessarily confusing. How does chemical biology differentiate itself from these established fields? The simple answer is that it does not. Each of the aforementioned disciplines arose when a few talented and farsighted chemists pivoted from contemporary problems in chemistry (which, as we are taught in high school, attempts to explain the properties of matter by understanding its structure and reactivity at an atomic level) to emerging challenges in biology.

Biochemistry seeks to reconstitute the essence of a biological phenomenon by placing a well-defined set of molecules in a highly controlled environment such as a test tube. Not only does biochemistry play a critical role in elucidating cell function, it also facilitates deeper insight into the chemistry of life. *Structural biology* elucidates the structures of spectacularly

doi:10.1162/DAED_a_00304

complex biological molecules and assemblies at an atomic level. To do so, it employs experimental and computational tools developed by chemists in the context of studying simpler forms of matter. The pioneers of *molecular biology* harnessed their insights into DNA structure and reactivity to transform biology from an observational science into an interventional one. Finally, *medicinal chemistry* emerged as a discipline when chemists were tasked with the goal of engineering potent drugs that modulated human physiology in a targeted manner.

In its broadest definition, chemical biology exploits a chemist's knowledge of molecular structure and reactivity, together with his or her skills in molecular design, synthesis, and analysis, to understand or engineer living organisms. In essence, this represents a return to the pioneering spirit of biochemists, structural biologists, molecular biologists, and medicinal chemists from an earlier generation. The difference today is that chemistry itself has become a far more powerful science than it was half a century ago. Our knowledge of the chemical properties of large swaths of the periodic table has grown enormously. This in turn has paved the road to cost-effective syntheses of incredibly complex molecules, some of which have become life-saving drugs. Similarly, back when antibiotics were first isolated from soil microbes, their biosynthetic origins were incomprehensible. Today, we can not only decode the chemical logic of antibiotic biosynthesis, but also engineer antibiotics ourselves. Meanwhile, chemistry's analytical methods have become so sophisticated that single molecules can be visualized with microscopes and useful information can be extracted from even the tiniest sample, such as a biopsy from a patient with an undefined illness. In this way, chemical biologists harness the evolving science of chemistry to interrogate or modify biology.

The interface between chemistry and human biology holds considerable promise for science, medicine, and society. At a fundamental level, chemistry can strengthen the foundations of human biology in a manner that is entirely analogous to its enabling role in many of biology's most notable advances in the twentieth century. Human beings are different from other living creatures (and indeed, even from each other) in many interesting ways: our brains, our diets, and our immune systems are just a few examples. Explaining these differences in the language of chemistry is a scientific frontier that has proven to have profound consequences for human health. Imagine a future where it is possible to meaningfully discuss the chemical basis of specific thoughts and emotions; or a time when our understanding of the immune system is deep enough to interpret its constantly changing responses to the effects of aging, diet, and infection on each of us. The age-old debate of nature versus nurture has taken on a new meaning with the discovery of epigenetic phenomena that can be passed on from one generation to the next. Unraveling this epigenetic code represents an exciting opportunity for chemical biologists to penetrate the mysteries of complex illnesses.

The chemistry–human biology interface also holds a special place in the future of drug discovery, development, and evaluation. As human beings struggle to reconcile their dreams for healthy aging with the need for cost-effective healthcare, innovative medicines are expected to be the panacea. A surprisingly large fraction of efforts to translate groundbreaking biological discoveries into patient care are bottlenecked by the lack of suitably engineered molecules or molecular assemblies. By focusing on problems where innovative molecular design, synthesis, or analysis is crucial, chemistry can accelerate the translation of advances in human biology into clinical practice. Take, for example, the field of infectious diseases. At a time when our armamentarium of effective antibiotics has reached alarmingly low levels, our knowledge of nature's repertoire of antibiotic biosynthetic strategies is exploding. Chemical biology is poised to exploit these insights to engineer pathogen-specific therapies. Consider also the field of radiology. New chemical probes and measurement methods such as MRI (magnetic resonance imaging), PET (positron emission tomography), and ultrasound have the potential to noninvasively visualize human anatomy and physiology effortlessly and at unprecedented resolution. Equipped with an unimaginably sophisticated set of accessories and apps, the iPhone of the future may be just as important a communication tool between healthcare consumers and providers as its present-day version is between two teenagers. Chemical biologists are also opening new doors in preventive medicine. Until recently, vaccines were principally used to protect against deadly or debilitating infectious diseases. Today, synthetic vaccines are being developed against cancer and allergy. Our bodies also play host to innumerable bacterial cells (collectively referred to as the "microbiome") whose myriad health benefits remain to be understood and perhaps even engineered. Last but not least, regulatory science represents an attractive but overlooked area for applied chemical biological research. By upgrading the capacity for risk-benefit analysis at agencies such as the FDA, innovative medicines could be brought to patients who need them the most, more cheaply and quickly than is currently possible.

Perhaps the most far-reaching impact of the convergence between chemistry and human biology will be at a pedagogical level. Recent years have witnessed a gradual de-emphasizing of organic chemistry in pre-medical education. While it is difficult to envision a resurgence of interest in chemistry in mainstream medical education, dual strengths in chemistry and medicine could foster a new breed of physician-scientists who are also physical scientists. Their talent for molecular design and analysis, coupled with their passion for human biology, would allow them to play leading roles in reshaping the healthcare industry.

Several compelling examples point toward the opportunity that lies ahead at the chemistry–human biology interface. A recent report on innovation in drug discovery, development, and evaluation by the President's Council of Advisors on Science and Technology highlighted both the continuing need for innovative medicines as well as widespread concerns about their development pace and cost.[1] The report proposed doubling the current annual output of innovative new medicines as an ambitious goal. While most industry watchers will concur that seventy to eighty innovative new drugs per year would indeed represent a dramatic increase in research and development productivity, two other figures put the importance of this goal into clearer perspective. First, of the roughly 23,000 proteins encoded by the human genome (the functions of as many as half of which remain unknown), fewer than 3 percent are targeted by FDA-approved drugs. Targeting proteins in the human body is by far the most productive approach to drug design. Second, there exist at least 10,000 diseases identified by International Classification of Diseases (ICD) – including more than 6,500 "orphan conditions" (disorders, often rare, for which drug development is commercially nonviable) – which lack effective therapies. Together, these numbers suggest that, at the present pace of drug discovery and development, the road to even a moderately comprehensive arsenal of human-grade drug treatments may be a long one. More fundamentally, in sharp contrast to biological studies on virtually every other model organism, human biology remains predominantly an observational science. Given that genetic manipulation of human beings is likely to remain severely constrained on ethical grounds, the nexus between chemistry and human biology needs to be strengthened in order for human biology to advance from its present status as a principally observational science into an interventional one.

Another argument supporting a serious re-evaluation of the chemistry–human biology interface is the current state of the pharmaceutical industry. The high cost of bringing a new drug to market (which, by some accounts, can run as high as $500 million) has forced the industry to prioritize discovery of the highest-priced medicines over new paradigms for affordable healthcare. The time is ripe for the emergence of new ideas and technologies that could spawn complementary business models and public-private partnerships to restructure the healthcare industry. The emergence of a thriving molecular biomarker industry is one example. Chemical biology could foster other analogous opportunities in the not-too-distant future. For example, innovations in molecular toxicology will be pivotal to the success of personalized medicine. Similarly, companies that harness the creativity of both chemistry and human immunology to develop fundamentally new approaches to preventive medicine will likely blaze new trails not too different from those forged by the pioneers of the Internet era.

Meanwhile, two technological advances – genome sequencing and stem cell culture – are not only propelling the emergence of *Homo sapiens* as one of biology's most attractive targets of investigation but are also making a strong case for a closer, intellectually deeper alliance between chemistry and human biology.

The first relates to the ease of sequencing the entire genome of a human being. The National Human Genome Research Institute estimates that the present cost of sequencing an individual's genome is under $10,000; this number is expected to fall by at least an order of magnitude in the foreseeable future. It is therefore very likely that, before too long, every healthcare consumer will be able to have his or her entire genome sequenced for the price of an MRI. The question now becomes: how can one exploit this information to enhance disease management or, better yet, healthy living? Cost-effective decoding of this data may well be one of chemical biology's greatest contributions to our society since the sequencing of the genome.

Biologists have independently developed ways to produce induced pluripotent stem cells (iPSCs): cells derived from adult humans that have been genetically reprogrammed into an embryonic stem cell–like state. Notwithstanding the infancy of this technology, iPSCs have the potential to revolutionize human biology, and leading medical centers are rushing in to establish facilities for routine generation of patient-derived iPSCs. The ability to use stem cells as individualized test tubes to understand, prevent, or treat disease represents an extremely promising opportunity for chemical biology.

Although its promise is clear, the field of chemical biology has not, to date, reached its full potential. Much effort focuses on harnessing robust chemistry in conjunction with high-speed engineering platforms in order to quickly translate emerging biological knowledge into new chemical tools. This is important work from which new drugs will inevitably emerge.[2] However, the real promise of chemical biology lies in two other pursuits. At a fundamental level, chemical biology can help elucidate what causes derangement of human physiology in the first place. There is growing consensus that complex diseases such as autism and autoimmunity are caused by genetic as well as environmental factors. If so, chemical biology may be able to shine light on the interplay between these triggers. And at the technological level, radically new molecular tool-making approaches are needed to transform knowledge of the biology of seemingly intractable diseases into practical treatments. Consider cystic fibrosis and von Gierke's disease (a glycogen storage disorder): the mutations responsible for these debilitating conditions were identified more than twenty years ago. Yet today we are no closer to translating these genetic insights into cures or even good treatments. The defective molecules in cystic fibrosis and von Gierke's disease are just two examples of scores of clinically relevant targets categorized as "undruggable" by today's drug discoverers. The human chemical biologist, on the other hand, recognizes that the reason why these genetic discoveries have fallen short is because the right kind of molecular tool has not yet been invented – but it can be done.

Several prominent academic institutions (including my own, Stanford University) have launched major initiatives at the chemistry-biology interface within the past decade. In most cases, these programs are collaborative efforts between existing chemical and biological science departments within the institution, although a few examples of cross-institutional efforts have also gained momentum. The ideal environment would bring together clinicians, scientists, and engineers, all of whom share an interest in strengthening the chemical foundations of human biology. These scholars will likely be either gifted molecular scientists or engineers who have turned their attention to important challenges in human biology, or they will be insightful biologists or physicians who can frame human biology's most fascinating mysteries in a manner that lends itself to chemical analysis or engineering. Bringing these researchers together is essential in order to encourage cross-fertilization of ideas and cultures. Environments where such collaborations occur will inevitably emerge as spectacularly powerful training grounds for a new breed of young "physician-scientist-engineers." These researchers will speak about human biology in the language of chemistry, tinker with objects on a length-scale one million times smaller than the width of a human hair, and only show deference to the laws of thermodynamics. Their talent for molecular toolmaking, coupled with their passion for human biology, will allow them to evolve new industries and business models where health, not sickness, drives the bottom line.

Each fall, I find myself facing a class of around two hundred of Stanford's most accomplished undergraduates taking their first course in biochemistry. I begin my first lecture with the reminder that, up until today, my students' chemical and biological educations have been orchestrated from separate universes, but this is about to change. After all, biology is chemistry, and chemistry would be not nearly as interesting were it not for biology. Many of my students go on to become successful doctors, scientists, or engineers. I can only wonder whether any of my students will someday return to Stanford as clinician-scientist-engineers. If so, what problems will they see that none of us do? And will they see fundamentally new solutions to problems that the rest of us consider unsolvable? I look forward to that day.

ENDNOTES

[1] President's Council of Advisors on Science and Technology, "Report to the President on Propelling Innovation in Drug Discovery, Development, and Evaluation," Executive Office of the President of the United States of America, September 2012.

[2] For a review of academic drug discovery operations, see Julie Frearson and Paul Wyatt, "Drug Discovery in Academia – The Third Way?" *Expert Opinion on Drug Discovery* 5 (2010): 909–919.

Using Computational Chemistry to Understand & Discover Chemical Reactions

K. N. Houk & Peng Liu

Abstract: Chemistry, the "science of matter," is the investigation of the fabulously complex interchanges of atoms and bonds that happen constantly throughout our universe and within all living things. Computational chemistry is the computer modeling of chemistry using mathematical equations that come from physics. The field was made possible by advances in computer algorithms and computer power and continues to flourish in step with developments in those areas. Computational chemistry can be thought of as both a time-lapse video that slows down processes by a quadrillion-fold and an ultramicroscope that provides a billion-fold magnification. Computational chemists can quantitatively simulate simple chemistry, such as the chemical reactions between molecules in interstellar space. The chemistry inside a living organism is dramatically more complicated and cannot be simulated exactly, but even here computational chemistry enables understanding and leads to discovery of previously unrecognized phenomena. This essay describes how computational chemistry has evolved into a potent force for progress in chemistry in the twenty-first century.

K. N. HOUK, a Fellow of the American Academy since 2002, is the Saul Winstein Chair in Organic Chemistry in the Department of Chemistry and Biochemistry at the University of California, Los Angeles.

PENG LIU is an Assistant Professor of Chemistry at the University of Pittsburgh.

(*See endnotes for complete contributor biographies.)

In chemistry class, we learn that chemists study matter and its properties; they wear lab coats and safety glasses and mix chemicals together and observe the amazing things that happen. But there is no need to go into a chemical laboratory to find chemistry. In fact, chemistry is literally everywhere: it is the thousands of chemical processes that result in the emergence of a growing plant from a seed, the transformation of flower nectar into the flight of a hummingbird, or the conversion in chemical factories of oil from decayed ancient life into polymers that are made into stylish fabrics or spacesuits. How do these things happen? Chemists learn how chemical reactions occur and how to control them for human purposes. In the twenty-first century, computational chemistry plays a major role in chemical discovery.

Before the twentieth century, knowledge about the properties and transformations of matter was gained

doi:10.1162/DAED_a_00305

through experimentation. Early chemical theories and rules, such as Mendeleev's periodic table, were empirically derived from observations of chemical phenomena. Some theories were wrong (for example, the phlogiston theory, which posited the existence of an element called phlogiston in order to explain combustion), while others were very crude models. The discovery of quantum mechanics in the 1920s revolutionized science. Heisenberg, Schrödinger, Dirac, and other physicists developed a theory based on pure mathematics that explains how chemistry arises from the interactions of nuclei and electrons.[1] Paul Dirac, one of the Nobel Laureates for quantum mechanics, noted in 1929:

> The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.[2]

Exactly as Dirac envisioned, a hierarchy of mathematical models, with different levels of approximation, has been developed over the last century.[3] But Dirac could not foresee the discovery and development of powerful computers with which we can solve some of these highly complex problems of applied mathematics. While we still cannot obtain exact solutions to the quantum mechanical equations for chemical systems with very large numbers of atoms, we can calculate answers as close as desired to the exact mathematical solution, given enough computer time. When more powerful computers become available, computational chemists will set out to solve bigger and bigger problems and try to obtain ever more accurate solutions to small problems.

Experiments yield facts, such as which products are formed when various chemicals come into contact or how much electricity is generated when sunlight shines on a chunk of silicon or sandwich of organic polymers. However, experiments do not tell us *why* such results occur. For example, why are certain products formed and not others, or why is only a few percent of the energy in sunlight converted to electricity? Both theory and computation are needed to answer these questions: theory to provide the general framework and simple models for a qualitative conceptual underpinning of experimental phenomena, and computation to flesh out an accurate microscopic account of them. Today's chemists attempt to employ computations to explain phenomena and guide new experiments, but quantitative modeling of chemical reactions is very challenging due to problems of scale. The chemical phenomena that we observe are the outcomes of rearrangements of the atomic structures of a huge number of very small molecules. A water droplet contains around one sextillion (10^{21}) molecules, each with a slightly different shape, velocity, and energy at any given moment. The atoms in each water molecule are moving rapidly inside the droplet: the atoms change to a new arrangement 10^{14} times per second. To completely reproduce the properties of that droplet and predict how it will change upon heating or mixing with other chemicals would require simulating all sextillion of the droplet's fast-moving molecules, were we to compute everything from exact quantum mechanical equations (or "first principles"). Modern computers can calculate how one molecule changes over time, but to calculate all sextillion or even a significant fraction of them is not practical, nor will it be anytime in the foreseeable future. However, approximate equations – model systems calibrated with empirical data to capture the average properties of a water molecule and its interactions with other molecules – can be computed to allow us to understand what occurs in the drop of water and to estimate its properties: density, surface tension, viscosity, and even chemical reactivity.

Aside from the daunting numbers of calculations that must be performed to mimic reality, there is also the issue of the size of some important molecules: smaller molecules are, of course, much simpler to model. As the number of electrons in a molecule increases, the time needed to perform calculations on it goes up rapidly. A hydrogen molecule (H_2) consists of two of the lightest atoms bonded together and only two electrons; natural gas consists primarily of methane (CH_4), which has only five light atoms and ten electrons. Everything about individual hydrogen and methane molecules can be computed nearly exactly in a short time. However, many molecules of crucial importance for life, and those that make up common materials, are much larger. Consider a nucleic acid molecule, such as a strand of DNA, or a protein that controls so many of the processes of life, or a polymer molecule in a polystyrene cup: each molecule contains thousands of atoms and can exist in many different three-dimensional arrangements that interconvert very quickly. To simulate the behavior of chemicals with so many atoms takes many computer resources. Depending on how accurate the calculations need to be, the number of hours needed to perform computations on molecules scale between the third and the seventh power of the number of electrons contained within them! This means, for example, a calculation involving a benzylpenicillin molecule with forty-one atoms can be up to two million times slower than the same calculation done with a methane molecule. Because of this, calculations on such large molecules must involve shortcuts that make the calculations faster but less accurate.

Computational modeling is the simulation of chemical structures, properties, and reactions with a computer. Simulation is sometimes described as the third form of science. The first form, experimental science, starts with empirical observations and models created from inductive logic. The second form is theoretical science, formulated in equations that describe the phenomena of the natural world. Simulation is a third form where mathematical equations are coded into computer programs to predict what happens in various hypothetical chemical situations.

The fundamental theories used in these computer programs are based on classical and quantum mechanics. Galileo, Kepler, Newton, and other scientific revolutionaries of the late seventeenth century developed what we now call classical mechanics, which describes the physics of relatively large objects moving on a human timescale. Newton's equations of motion (as these classical mechanics equations are often called) are used for molecular dynamics simulations to derive the motion of atoms in molecules or larger objects over time. Classical mechanics can also be used to study structures of molecules by fitting equations to empirical data – what chemists call "molecular mechanics." However, classical mechanics cannot predict chemical reactions and reactivity, because the motions of electrons are wavelike, quantized, and described correctly only by quantum mechanics. For more massive and slowly moving systems, classical and quantum mechanics converge, but quantum mechanics is uniquely capable of describing the electronic structure of atoms and molecules and thus chemical properties and reactions. Multiscale computational methods, which employ both classical and quantum mechanics, have been developed for calculations of complex chemical and biological systems, such as proteins. Here, quantum mechanics is applied to study the central part of the system: for example, atoms that are close to the forming or breaking chemical bond in a reaction. The remaining atoms are treated with classical mechanics so that such calculations can be applied to very large systems. In 2013, three of the pioneers in this field, Martin Karplus, Michael Levitt, and Arieh Warshel, were awarded the Nobel Prize in Chemistry for their studies in the early 1970s that established what is now called the QM/MM method.

Based on these underlying theories, many computer algorithms to calculate properties and reactions were written in the last century, and these developments continue to this day. While the fundamental equations of quantum mechanics are deceptively simple, the computer programs written in order to use them in simulations are extremely complicated. In 1998, the Nobel Prize in Chemistry was awarded to John Pople, a mathematician and chemist whose research group developed many of these algorithms and computer programs, and Walter Kohn, a physicist who, with his coworkers, developed an alternative method of solving the Schrödinger equation, now known as density functional theory (DFT). Pople and Kohn were at the forefront of using computational methods in chemistry, inspiring many mathematicians, mathematical chemists, and physicists to devise algorithms and computer methods for studying chemical phenomena. For example, one of the main programs now used for these calculations has seventy-four authors from all over the world![4] These authors and many other scientists worked over the last fifty years on various computational chemistry programs that are now in general use by chemists. The programs have become very useful for exploring real chemistry, and software companies have been formed to further develop and market these programs commercially.[5]

The progress described here was stimulated by the development of computers. The ENIAC (Electronic Numerical Integrator and Computer) and other general-purpose computers in the 1940s occupied space equal to a comfortable house for four people. Now, the computers that reside in our smartphones are about a trillion times more powerful. Furthermore, computational chemists have access to computers all over the world, and rapid Internet connections give computational research groups in the United States access to a whole network of powerful computers supported by the National Science Foundation and other federal research agencies. Other countries have similar networks, and there is international competition to produce the most powerful computer.

The impact of these developments on the capabilities of computational chemistry has been profound, and the field has become an increasingly important aspect of science. In the flagship journal of chemistry, the *Journal of the American Chemical Society*, the number of computational papers has risen from very few in the 1960s to over three hundred papers per year. Along with the growth of computational chemistry in mainstream journals, there has been a proliferation of journals specifically devoted to the subject. There is the *Journal of Computational Chemistry*, the *Journal of Chemical Theory and Computation*, and at least two dozen other journals that concentrate on the study of chemistry using computation. Chemistry is not unique in this regard: physics has a dozen such journals; and biology has *Computational Biology* and many other publications that emphasize computation.

The most successful computations also lead to the development of general concepts that can be used to guide future experiments and make predictions. This is, of course, helpful in the field of organic chemistry, the area of expertise of the authors of this article. Organic chemistry may be notorious as a gatekeeper for future doctors, but it is really an intellectually rich and challenging branch of chemistry that involves chemical compounds containing carbon atoms, along with any of the other atoms of the periodic table. Organic chemistry touches all of our existence, from lifesaving and -enhancing pharmaceuticals to fuels, insecticides, and organic electronic materials. We describe in the following pages how computations are used to explore and understand organic chemistry.

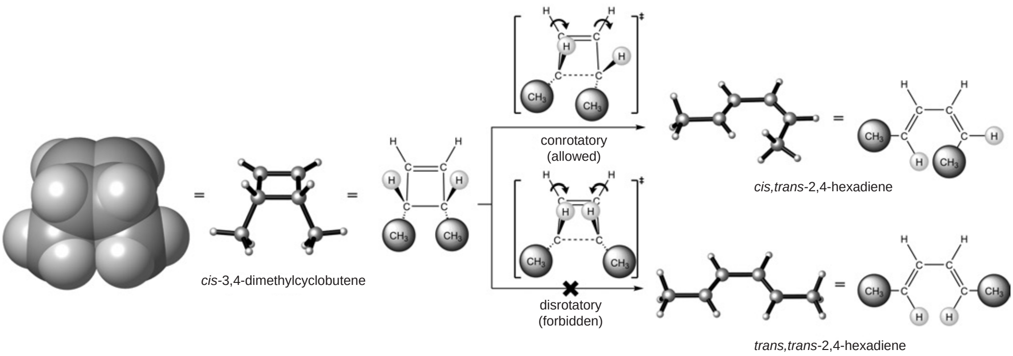
In 1965, R. B. Woodward and Roald Hoffmann published one of the most influential conceptual developments that thrust theory, and eventually computation, into the forefront of organic chemistry.[6] Although these concepts were grounded in previous developments by many scientists, they came to be known by chemists as the Woodward-Hoffmann rules. Based on quantum mechanical principles, these rules give predictions about a particular class of organic chemical reactions in which bond formation and bond breakage occur simultaneously in a ring of atoms. One example of such a reaction is the ring opening of a molecule known as *cis*-3,4-dimethylcyclobutene, shown in Figure 1.

With these images, we launch into real organic chemistry and hope to introduce the reader to the visual world that organic chemists occupy and that computational organic chemists study. The first two pictures shown in Figure 1 are computer drawings of the three-dimensional structure of the *cis*-3,4-dimethylcyclobutene molecule. While it can also be represented by its formula, C_6H_{10}, there are many different molecules with that same formula, each of which has unique properties. The exact way that the atoms are arranged in this figure was predicted by a quantum mechanical calculation. The calculation took ten minutes today using a powerful desktop computer, but thirty years ago when the study began, the same calculations took one week and involved a small roomful of equipment. Chemists have developed these types of pictures for the rapid visual representation of what is actually a very complicated mathematical result in a computer.

Experiments already showed that these reactions typically occur with the rotation of both termini in the same direction (this motion was called "conrotatory" by Woodward and Hoffmann), rather than opposite directions (called "disrotatory"). As shown in Figure 1, each of the two motions leads to a distinct product called a stereoisomer. Different stereoisomers have the same atoms and bonds, but they are connected together in different three-dimensional spatial arrangements. This difference in shape gives stereoisomers distinct chemical properties: molecules that have the same "structure" but a different stereochemistry and shape may turn out to be a life-saving drug or a poison, depending on their three-dimensional shape. It is therefore important to understand and control which stereoisomer is formed in a reaction used to make the molecule.

Using qualitative reasoning and supported by the very approximate calculations possible at the time, Woodward and Hoffmann provided an elegant quantum mechanical interpretation of the selectivity shown in Figure 1. The conrotatory process maximizes bonding all along the reaction pathway (it is "allowed," or occurs rapidly), while the disrotatory opening involves a motion that would require a very high energy to occur (it is therefore "forbidden"). They described these principles in terms of the symmetries of orbitals, the regions in which electrons are located according to the usual form of quantum mechanics used to describe molecules. The insights led to

Figure 1
Three-Dimensional "Space-Filling," "Ball-and-Stick," and Schematic Representations of *cis*-3,4-dimethylcyclobutene and the Conrotatory and Disrotatory Reaction Products


=
=
cis-3,4-dimethylcyclobutene
H
C
CH_3
conrotatory
(allowed)
disrotatory
(forbidden)
‡
=
cis,trans-2,4-hexadiene
trans,trans-2,4-hexadiene

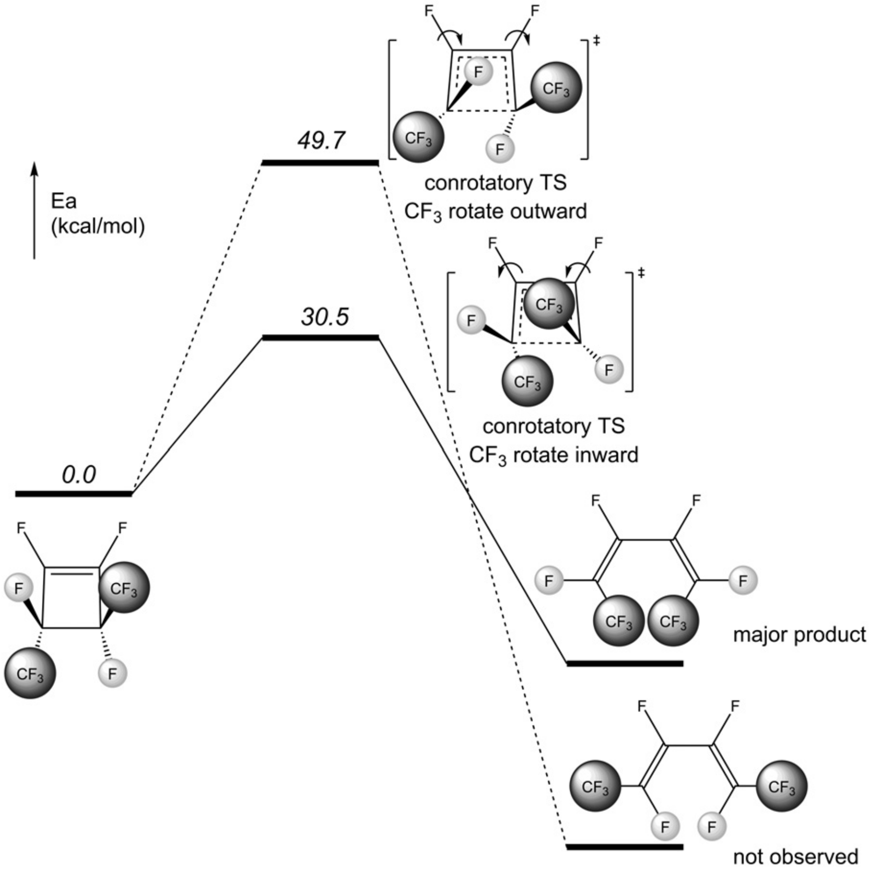
In the two structures on the far left, spheres represent the positions of the atoms in the reactant molecule. The larger spheres show the carbon atoms, and the smaller spheres show the hydrogen atoms. The size of the atoms in the "space-filling" picture represents their van der Waals radii (which measure how close two atoms can approach). Smaller spheres were used for the "ball-and-stick" picture to illustrate the chemical bonds (indicated by bold lines) that are formed by a buildup of electrons between the nuclei. The third picture is a sketch of the same molecule, with the atoms represented by letters and the large balls representing the larger size of the "substituent" methyl groups (CH_3; the bold lines indicate that the H atoms are in the foreground, and the dotted lines indicate that the CH_3 groups recede backward). The sketches in brackets show the changes occurring in the reactions. The dashed lines indicate bonds that are breaking in the reaction. The products of the reactions are shown in the "ball-and-stick" and sketch renditions to the right of the arrows. Source: Figure prepared by the authors using data in R. Hoffmann and R. B. Woodward, "Stereochemistry of Electrocyclic Reactions," *Journal of the American Chemical Society* 87 (1965): 395–397.

new understanding of a broad segment of organic chemistry and to predictions of new reactions that were subsequently discovered experimentally.

Chemists found through experiments that their understanding was incomplete. In examples such as that shown in Figure 2, there are two different "allowed" conrotatory processes: namely, rotation in a clockwise or in a counterclockwise fashion. The Woodward-Hoffmann rules did not differentiate between the two, but researchers found a huge preference for one direction of rotation in several cases studied experimentally.[7] The difference between the activation energies required for these processes to occur (30.5 kcal/mol for the inward conrotatory rotation and 49.7 kcal/mol for the outward conrotatory rotation) is enough to make the observed inward-rotating reaction ten billion times faster than the non-observed outward-rotating reaction. The result was very puzzling, since the larger CF_3 group bumps into the other CF_3 in the faster reaction, and nature usually minimizes such bumping (which we call steric clashes) – but not here. Our group used quantum mechanics and computational chemistry to understand why.

Quantum mechanical simulations of these reactions showed the motions of the nuclei and electrons in these molecules as they change from reactants to either of the two possible products. These calculations also determine how much energy it takes for the bonds of one molecule (the reactant) to be reorganized through nuclear and electron rearrangements to form another molecule (the product). The "transi-

Figure 2
Two "Allowed" Reactions and the Large Activation Energy Differences in the Two Modes of Ring Opening of *trans*-perfluoro-3,4-dimethylcyclobutene



The horizontal lines represent the relative energies of the reactants (left), transition states ("TS," center), and products (right), and the lines show how these are interrelated. The experimentally measured activation energies (E_a) are shown next to the corresponding transition states. Lower activation energy means a faster reaction, and so the reaction follows the path with the lowest activation energy (indicated by the solid lines), even though the product of that path is less stable. Source: Figure prepared by Peng Liu using data from W. R. Dolbier, Jr., H. Koroniak, D. J. Burton, A. R. Bailey, G. S. Shaw, and S. W. Hansen, "Remarkable, Contrasteric, Electrocyclic Ring Opening of a Cyclobutene," *Journal of the American Chemical Society* 106 (1984): 1871–1872.

tion state" is the highest-energy point along the best path from reactant to product as marked in Figure 2,[8] and it is the energy of this transition state relative to the reactants (activation energy, E_a) that determines how fast or slow the reaction will occur. By quantum mechanical calculations, we found out what the transition states look like (rough sketches are given in Figure 2) so that we could analyze them to understand why one transition state is much lower in energy than the other.

Calculations of this type could also be performed for other atoms and groups besides F and CF_3, and we eventually learned that certain types of substituents – those we call electron-donors (D in Figure 3) – always rotate outward away from the breaking bond, but strong electron-acceptors (A in Figure 3) rotate inward toward the breaking bond. Donors are already surrounded by electrons and thus avoid interacting with the electrons of the bond that breaks, but acceptors seek electrons and tend to move toward the breaking bond (see Figure 3). The shaded shapes are meant to represent the regions where electrons are localized and would have repulsive interactions with other filled orbitals nearby. The empty shapes represent regions that do not have electrons but would like to; these empty orbitals cause an atom to be electron-loving, or "electrophilic." Quantum mechanics shows that the interaction of a filled orbital with a vacant orbital is favorable, while the interaction of two filled orbitals is repulsive. This is the basis of bonding and steric effects, respectively.

Since the bond twists – or torques – as it breaks, we describe this selective twisting as "torquoselectivity."[9] Computations were used first to reproduce the initial experiment, then to analyze the results, and then to develop concepts and rules to predict the results of similar future experiments. The principle here has been applied to predict the course of reactions of new substances synthesized for the first time. For example, Figure 4 shows a simple molecule, 3-formylcyclobutene, which was made in our laboratory in 1987 for the first time to test the prediction made about the unexpected stereochemistry of this reaction.[10] Based upon our torquoselectivity theory, we expected that the formyl group (labeled "CHO" in Figure 4), an acceptor-type substituent, would rotate inward, and we used a quantum mechanical simulation to predict exactly how much this was preferred over an outward rotation. Then we did the experiment, and it worked! Only the less stable product (shown on the top line of Figure 4) is formed.

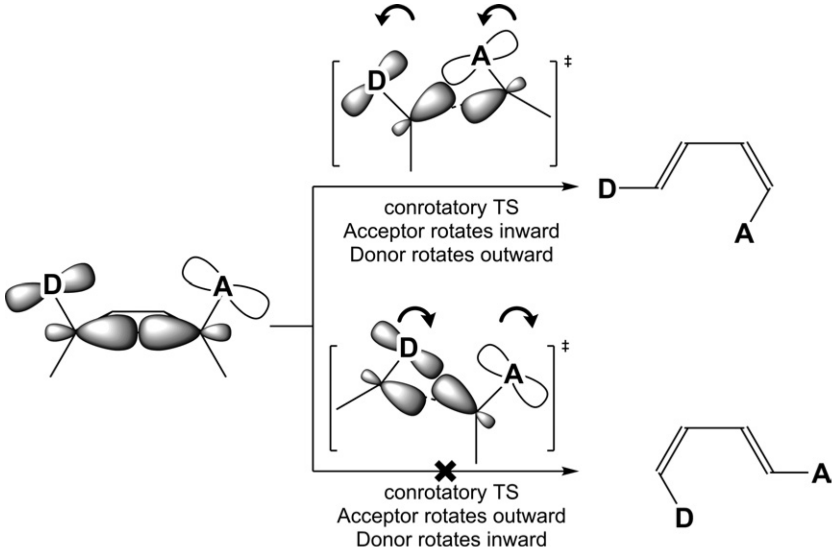
This example illustrates that when quantum mechanics is applied to a small enough molecule, an accurate prediction of the product of a new chemical reaction is possible. These calculations also led to the development of the theory of torquoselectivity that is applicable to every other reaction of this type.

Another significant application for computational chemistry is the development of catalysts (substances that speed up a reaction but are not consumed during the reaction). Catalysts cause reactions that normally do not occur at all to take place under conditions that are easily achievable. Chemists aim to develop catalysts to achieve new chemical transformations or to increase the efficiency of valuable reactions.

Many reactions used in the chemical industry involve transition metal catalysts (transition metals are so called because they have partially occupied *d* orbitals and therefore special properties). Three Nobel Prizes in Chemistry have been awarded in the last fifteen years (in 2001, 2005, and 2010) for discoveries about organic reactions using these catalysts.[11] The study of chemicals containing carbon and metal atoms in the same molecule is called organometallic chemistry and is now a prominent field of chemistry. The carbon atoms are part of the ligand attached to the metal.

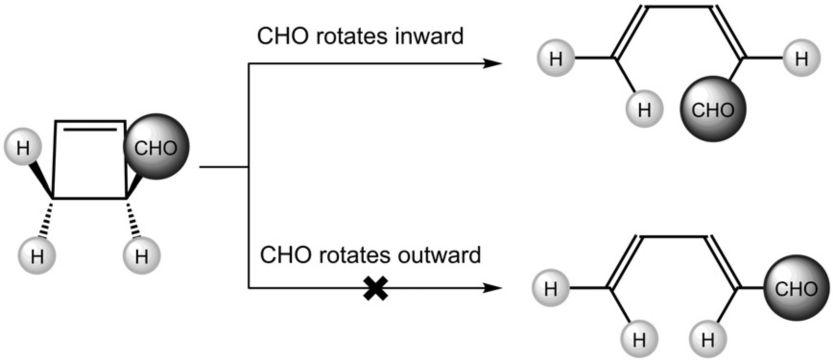
Catalytic reactions generally involve many steps and intermediates that are usually difficult to detect or identify experimentally, although new spectroscopic and imaging tools are being developed to try to achieve this. In the glory days of mechanistic physical organic chemistry, the masters of the field, such as Saul Winstein at UCLA and Paul D. Bartlett at Harvard, devised many clever experiments using the instrumentation available at the time to try to deduce how reactions occurred in solutions. Major controversies often devel-

Figure 3
Orbital Interactions in Two Conrotatory Transition States in the Electrocyclic Ring Opening of a *cis*-3-donor-4-acceptor-cyclobutene



The torquoselectivity model developed from calculations predicts (correctly) that the counterclockwise motion shown on the top line of the figure is highly preferred. Source: Figure prepared by the authors using data from N. G. Rondan and K. N. Houk, "Theory of Stereoselection in Conrotatory Electrocyclic Reactions of Substituted Cyclobutenes," *Journal of the American Chemical Society* 107 (1985): 2099–2111.

Figure 4
Computations Predicted the Torquoselectivity of Electrocyclic Ring Opening of 3-Formylcyclobutene



Source: Figure prepared by the authors using data from K. Rudolf, D. C. Spellmeyer, and K. N. Houk, "Prediction and Experimental Verification of the Stereoselective Electrocyclization of 3-Formylcyclobutene," *Journal of Organic Chemistry* 52 (1987): 3708–3710.

oped about the interpretation of the experiments. Nowadays we try to gain this mechanistic information about catalytic reactions by looking directly at molecules as they react using computer simulations.

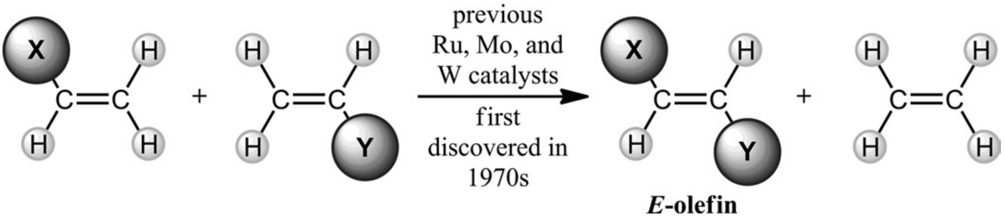
Shown in Figure 5 is olefin metathesis, a very important reaction in industry and laboratory synthesis. Metathesis means transposition, in this case of the atoms from one olefin to another. An olefin has two carbons joined by a double bond; each single bond is made of one pair of electrons (a double bond, comprising two pairs of electrons, is represented by two lines). The metathesis reaction shown in Figure 5 swaps the atoms making up the ends of each double-bonded olefin. This reaction provides one of the most powerful strategies for making new carbon-carbon double bonds and is important for the synthesis of many complex organic compounds and polymeric materials like those used for many familiar objects, from Norsorex pants for motorcyclists to gigantic wind turbine (modern windmill) blades.

A major challenge that developed during the study of olefin metathesis was to find catalysts that form "*Z*-olefins," in which the two substituents (X and Y) in the product are on the same side of the double bond. Many years after the discovery of olefin metathesis, chemists were still trying to learn how to make *Z*-olefins this way so that new compounds and materials could be synthesized via the olefin metathesis process. In 2009, chemists Amir Hoveyda and Richard R. Schrock found the first molybdenum- and tungsten-based olefin metathesis catalysts that selectively produce *Z*-olefins,[12] and two years later, chemist Robert H. Grubbs, one of the Nobel Laureates in this field (along with Richard R. Schrock and Yves Chauvin), discovered a new type of ruthenium catalyst that performed the same function (Figure 6). Why this catalyst produced *Z*-olefins, however, was not known.[13]

Determining why this ruthenium catalyst produced *Z*-olefins was a difficult challenge to computational chemistry. The molecules involved in this reaction are large and contain metals, which have high atomic numbers and dozens of electrons. Consequently, hundreds of computer hours are needed for each computation. In addition, there are many structures to compute due to the great number of ways the reaction could occur and the multiple structures involved in each case. Extensive experimental and computational studies of olefin metathesis with previously reported ruthenium catalysts have been carried out all over the world in the last few decades. Based on those studies, our group had many clues about what types of reactions we needed to investigate. We were able to limit the number of structures to evaluate, rather than having to compute every possibility for this complex reaction.[14] Given previous results, there were really only two plausible pathways, distinguished from one another by the direction from which the olefin molecule approaches the catalyst. The approach can be either adjacent or opposite to the ligand. These two pathways are shown in Figure 7 and are called "side" and "bottom" approaches.

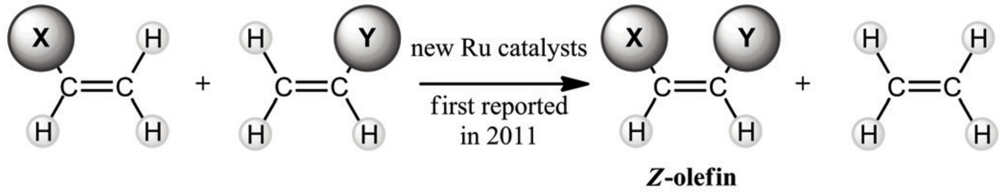
Our computations revealed a major surprise and a crucial discovery: in contrast to everything known before, this reaction with the new ruthenium catalysts involves side approach of the olefin to the catalyst.[15] Computational technology allowed us to render visualizations of the three-dimensional structures of the transition states (Figure 7). Scientists use microscopes to see microbes and employ atomic force microscopes (AFM) to study materials at the atomic level (for example, in the burgeoning field of nanochemistry). By contrast, there are currently no established experimental tools to visualize transition states, since they are only about 10^{-9} meters (1 nanometer) in diameter and exist for

Figure 5
The Olefin Metathesis Reaction Swaps the Ends of Olefins



The usual products of these reactions, called *E*-olefins, have the X and Y groups on opposite sides of the double bond. Source: Figure prepared by the authors using data from R. H. Grubbs and S. Chang, "Recent Advances in Olefin Metathesis and Its Application in Organic Synthesis," *Tetrahedron* 54 (1998): 4413 – 4450.

Figure 6
The Olefin Metathesis to Give *Z*-Olefins

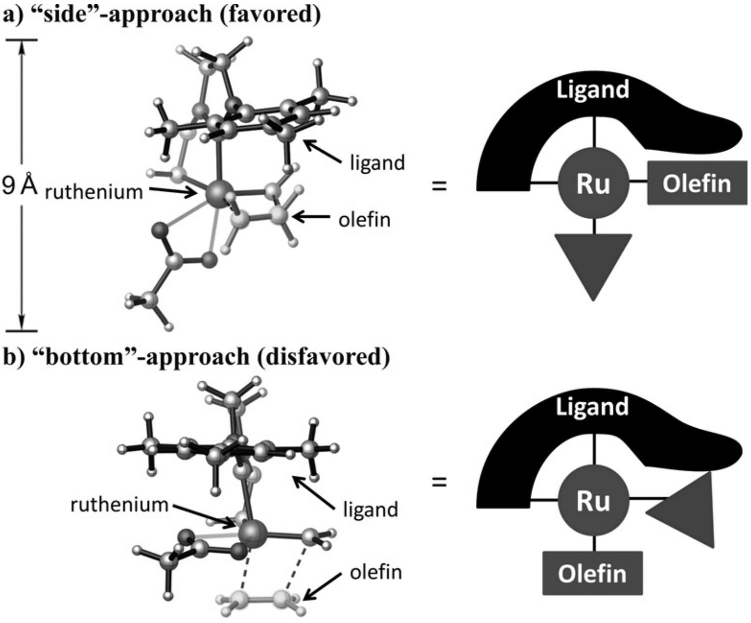


Source: Figure prepared by the authors using data from K. Endo and R. H. Grubbs, "Chelated Ruthenium Catalysts for Z-Selective Olefin Metathesis," *Journal of the American Chemical Society* 133 (2011): 8525 – 8527.

less than 10^{-13} seconds! (Chemists like Ahmed Zewail at Caltech are working to develop such experimental tools.) Computations, however, are able to bring the transition state to life by taking snapshots of simulated reactions as they happen (imagine a camera with a shutter speed of one femtosecond, or 10^{-15} seconds!) and by functioning like a super–high power microscope (with 10^9 times magnification!). Although it is a prediction that cannot currently be verified directly, this picture enables us to interpret important occurrences in this reaction, such as how individual atoms attract or repel each other, that would otherwise be impossible to observe. Such an analysis revealed that the *Z*-olefin is selectively formed due to the "side" approach of the olefin molecule in the catalyst complex, as shown in Figure 7. The olefin approaching in this way clashes with the ligand and places the substituents on the olefin on the same side of the newly formed double bond to form the *Z*-olefin. This is very different from previous reactions using other ruthenium catalysts, in which the olefin approaches from the bottom, far away from the ligand, causing the formation of more stable *E*-olefins rather than *Z*-olefins.

These computations provided important insights for further catalyst development. Armed with the knowledge that *Z*-selectivity in the new catalysts arises from the repulsions with the ligand on the catalyst, researchers began experimental studies of catalysts with even larger ligands. This led to the discovery of an improved *Z*-selective

Figure 7
Three-Dimensional Renditions of the Computed Transition State Structures of the Possible Approaches of the Olefin Molecule



a) Shows the olefin approaching the ruthenium catalyst adjacent to the ligand (the "side" approach); b) shows the olefin approaching opposite the ligand on the ruthenium catalyst (the "bottom" approach). A qualitative rendering is shown at the right. Source: Figure prepared by the authors using data from P. Liu, X. Xu, X. Dong, B. K. Keitz, M. B. Herbert, R. H. Grubbs, and K. N. Houk, "Z-Selectivity in Olefin Metathesis with Chelated Ru Catalysts: Computational Studies of Mechanism and Selectivity," *Journal of the American Chemical Society* 134 (2012): 1464–1467.

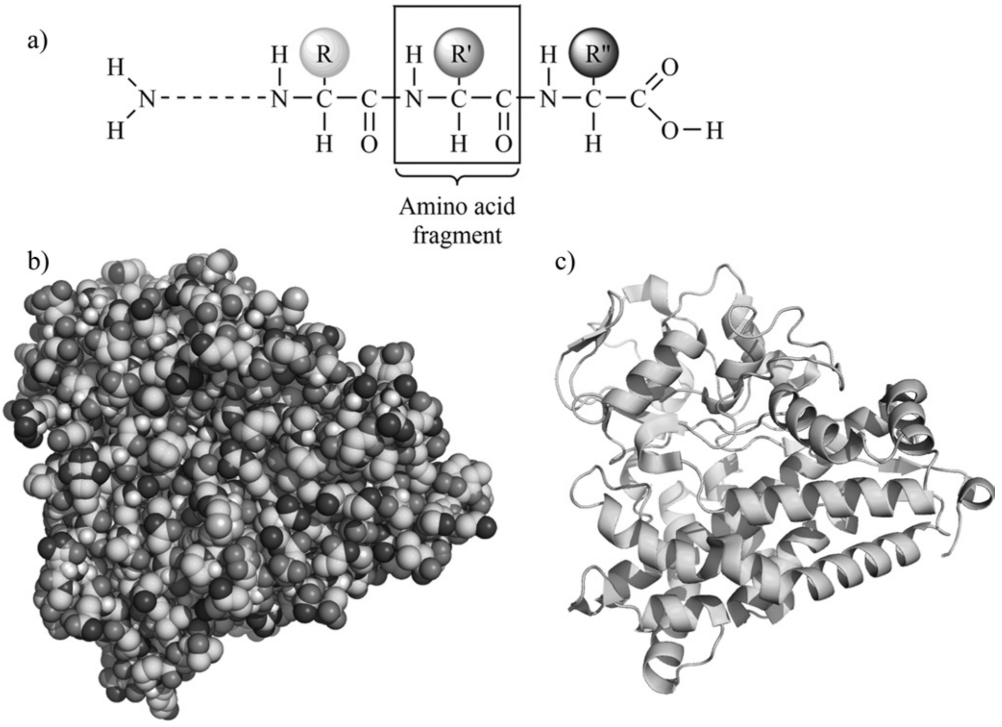
catalyst by the Grubbs group.[16] Computational investigations of this type have become a standard way to accelerate understanding and discovery, and many experimental groups have become involved in computational work to complement their experiments.

The *Z*-selective catalyst was discovered accidentally through experiments; now computations have helped to determine precisely which experiments might improve such catalysts. Similar computational approaches are also being used to predict new catalysts for pharmaceuticals, fuels, and materials. For example, computational materials scientists calculate how different combinations of metals produce useful metal alloys as catalysts.[17]

Nature generally uses proteins, and sometimes ribonucleic acids (RNA), to catalyze the reactions necessary for metabolism at the rates required to sustain life. Proteins are poly-amino acids of the general structure shown in Figure 8 (a), where "R" can be any of the twenty different side chains of natural amino acids. The amino acid fragments are connected in a specific sequence that determines the structure and properties of a protein.

Our research group at UCLA collaborates with David Baker's group at the University of Washington and Stephen Mayo's group at Caltech to design new enzymes

Figure 8
The General Structure or "Primary Sequence" of a Protein (a); the Three-Dimensional Representations of a Protein, Cytochrome P450cam, Showing All Atoms in a Space-filling Display (b); and a "Ribbon Diagram" of Protein Architecture (c).



Source: The three-dimensional protein structures are illustrated using PyMol (Version 1.3 Schrödinger, LLC) with crystal structure obtained from the Protein Data Bank (PDB ID: 2ZWT); and K. Sakurai, H. Shimada, T. Hayashi, and T. Tsukihara, "Substrate Binding Induces Structural Changes in Cytochrome P450cam," *Acta Crystallographica Section F* 65 (2009): 80 – 83.

that catalyze "non-natural reactions": the many reactions not catalyzed by naturally occurring enzymes. We do this by using computer calculations to predict which protein structures will fold into a specific three-dimensional structure like those shown in Figure 8 (b) and (c). We can then try to create a fold that will align the catalytic groups from the protein in order to catalyze a desired reaction – perhaps one that has known practical or commercial value or perhaps simply one we dreamed up. If we can predict the amino acid sequence (the list of individual amino acids and the order in which they are connected) needed, it is a simple matter for chemists to use automatic machines to synthesize the desired DNA and for molecular biologists to incorporate this DNA into a microorganism and induce it to produce these new proteins.

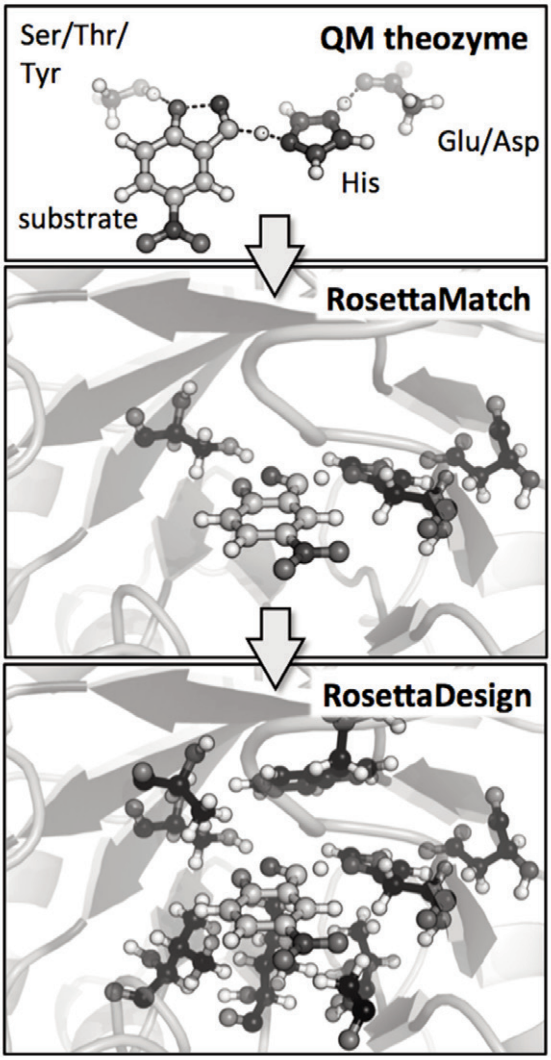
We use quantum mechanical calculations to design optimal arrangements of protein active site components that are predicted to catalyze specific reactions. David Baker's group has developed a computer program called Rosetta that predicts the amino acid sequences of proteins that will fold up into a specific three-dimensional structure. The program is based on classical mechanics and empirical information; because the proteins that are studied are large and can adopt many shapes, accurate quantum mechanical calculations would take too long to be useful. Although they only produce approximate models, Rosetta and other programs can nevertheless offer valuable information about which arrangement of atoms in proteins is most stable.

Realizing the potential of these tools, our group and David Baker's began about ten years ago to create what we call the inside-out approach to enzyme design (outlined in Figure 9).[18] We start with quantum mechanical models of the transition states of the chemical reaction we wish to catalyze and calculate which protein side chains will stabilize the transition state. This becomes a model for the core of the enzyme where catalytic reactions occur. We call this computational model a theoretical enzyme, or "theozyme." Then, using Rosetta, we find a stable protein structure in the Protein Data Bank (a database of the three-dimensional structures of all known proteins) that can be modified to achieve the designed structure with the necessary catalytic groups aligned in the perfect positions for catalysis.[19] After extensive computational tests using both quantum mechanics and classical molecular dynamics, the best computationally designed enzymes are selected for experimental testing. The actual proteins are produced by modified microorganisms such as *E. coli* and are then tested for catalysis. Using this procedure, we have successfully produced new effective catalytic proteins for three different types of reactions.[20] The entire process of designing new enzymes through computation currently takes years; however, it takes much longer (billions of years) for nature to evolve enzymes for metabolism. Right now we must do many thousands of calculations to make predictions of new enzymes, and even then only a small fraction of the computationally designed enzymes are active in the experiment. Nevertheless, designing protein catalysts from scratch using only computer calculations is a major development, and we envision that this technology will lead to catalysts for the synthesis of many important compounds and for therapeutic purposes as well.

These examples illustrate a very small portion of the field of computational chemistry. Computer programs that calculate and predict properties of chemical systems using a combination of theoretical methods have been developed for use in many areas of chemistry. One well-established example of the integration of multiple computational tools to solve important problems is the field of computational drug design.[21] Calculations in this enterprise range from structural evaluation to quickly screen thousands of candidate molecules for use in drugs to elaborate simulations of substrate-protein binding that can predict whether a molecule will act as a good inhibitor for a target protein involved in a disease. Such calculations have proven their worth in developing new enzyme inhibitors, although the path from effective inhibitors to commercial drugs is still long, expensive, and mostly empirical.

Yet another innovative use of computational chemistry is in developing new materials for many different industries. Computational chemists are developing programs based on a combination of quantum and classical mechanics to compute the properties of structural materials, solar-energy conversion devices, and new chemical batteries. Aiming to aid the development of computational architecture and methodology for materials chemistry, the White House approved the Materials Genome Initiative in 2011.[22] The name evokes the remarkably successful Human Genome

Figure 9
Overview of the "Inside-Out" Computational Approach to Designing Unnatural Enzymes



Source: G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, and K. N. Houk, "Computational Enzyme Design," *Angewandte Chemie International Edition* 52 (2013): 5700–5725.

Project and extends the idea to the world of materials. It has therefore been recognized at the highest policy level that computational methods can accelerate the discovery of advanced materials and shorten the process of deploying them to the commercial market.

Innovations in computer hardware design will continue to enhance the scope of computational chemistry. For example, the development of graphics processing units (GPUs) by the computer industry has energized the entertainment industry. The success of computer gaming has made these devices inexpensive, and computational chemists are rushing to adapt their programs to these commodity devices in order to enhance their modeling capabilities.[23]

What will we be able to do with these computers of the future? We have discussed in this article how computations are applied to study the way chemical reactions occur and to improve and extend them. In the future, this will be done more accurately, more quickly, and on much larger systems, producing more realistic models of chemistry. Predicting completely new chemical transformations is likely to remain challenging, because so many different bonds may be made or broken in a chemical reaction and, as we stated above, the combinations of even relatively simple pure chemicals can lead to a huge number of reaction pathways that grows exponentially with the number of atoms involved. Experimental knowledge about existing reactions may help chemists guess the outcome of an unknown reaction, but important discoveries in chemistry often result from discoveries of new types of transformations that occur in unexpected ways, not from simple extensions of known phenomena. Quantum mechanics can predict things that have never been observed. To predict what reaction happens when a new combination of chemicals is tested requires the evaluation of every possibility. Computational chemists are working on methods to predict reactions and their rates based solely on the information about the separated reactants, catalysts, solvents, and reaction conditions, essentially forcing molecules together and seeing what happens in the computer.[24]

We have described how computational chemists go about exploring chemistry and developing new models and theories to understand nature and predict useful things. Better algorithms and increasing computer power make ever larger and more accurate calculations possible, and this challenges computational chemists to take on larger and more complex problems. Computational chemistry has grown from the breakthrough theory of the early twentieth century into a ubiquitous and powerful engine for chemical discovery in the twenty-first.

ENDNOTES

[*] Contributor Biographies: K. N. HOUK, a Fellow of the American Academy since 2002, is the Saul Winstein Chair in Organic Chemistry in the Department of Chemistry and Biochemistry at the University of California, Los Angeles. He taught earlier at Louisiana State University and the University of Pittsburgh, and was Director of the Chemistry Division of the National Science Foundation. Over his career, he has "evolved" from an experimental physical organic chemist to a computational chemist, parallel to the developments of the research field described in this article.

PENG LIU is an Assistant Professor of Chemistry at the University of Pittsburgh. He obtained his Ph.D. in Chemistry at the University of California, Los Angeles, where he was a Postdoctoral Scholar in Professor K. N. Houk's research group. His research interests include computational studies of organometallic and organic reactions.

[1] Graham Farmelo, *The Strangest Man: The Hidden Life of Paul Dirac, Mystic of the Atom* (New York: Basic Books, 2009).

[2] Paul A. M. Dirac, "Quantum Mechanics of Many-Electron Systems," *Proceedings of the Royal Society of London A* 123 (1929): 714–733.

[3] Christopher J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd ed. (Malden, Mass.: John Wiley & Sons, 2004).

[4] M. J. Frisch et al., *Gaussian* 09, Revision D.01 [electronic structure modeling program] (Wallingford, Conn.: Gaussian, Inc., 2009).

[5] A. B. Richon, "An Early History of the Molecular Modeling Industry," *Drug Discovery Today* 13 (2008): 659–664.

[6] See R. Hoffmann and R. B. Woodward, "Stereochemistry of Electrocyclic Reactions," *Journal of the American Chemical Society* 87 (1965): 395–397; R. Hoffmann and R. B. Woodward, "Selection Rules for Concerted Cycloaddition Reactions," *Journal of the American Chemical Society* 87 (1965): 2046–2048; R. Hoffmann and R. B. Woodward, "Selection Rules for Sigmatropic Reactions," *Journal of the American Chemical Society* 87 (1965): 2511–2513; and R. B. Woodward and R. Hoffmann, *The Conservation of Orbital Symmetry* (New York: Academic Press, 1970).

[7] W. R. Dolbier, Jr., H. Koroniak, D. J. Burton, A. R. Bailey, G. S. Shaw, and S. W. Hansen, "Remarkable, Contrasteric, Electrocyclic Ring Opening of a Cyclobutene," *Journal of the American Chemical Society* 106 (1984): 1871–1872.

[8] Henry Eyring, "The Activated Complex in Chemical Reactions," *The Journal of Chemical Physics* 3 (1935): 107–115.

[9] W. Kirmse, N. G. Rondan, and K. N. Houk, "Stereoselective Substituent Effects on Conrotatory Electrocyclic Reactions of Cyclobutenes," *Journal of the American Chemical Society* 106 (1984): 7989–7991. See also N. G. Rondan and K. N. Houk, "Theory of Stereoselection in Conrotatory Electrocyclic Reactions of Substituted Cyclobutenes," *Journal of the American Chemical Society* 107 (1985): 2099–2111.

[10] K. Rudolf, D. C. Spellmeyer, and K. N. Houk, "Prediction and Experimental Verification of the Stereoselective Electrocyclization of 3-Formylcyclobutene," *The Journal of Organic Chemistry* 52 (1987): 3708–3710.

[11] These Nobel Prizes in Chemistry were awarded for asymmetric hydrogenations and oxidations (William S. Knowles, Ryoji Noyori, and K. Barry Sharpless; 2001), olefin metathesis (Yves Chauvin, Robert H. Grubbs, and Richard R. Schrock; 2005), and palladium-catalyzed cross couplings (Richard F. Heck, Ei-ichi Negishi, and Akira Suzuki; 2010).

[12] A. J. Jiang, Y. Zhao, R. R. Schrock, and A. H. Hoveyda, "Highly *Z*-Selective Metathesis Homocoupling of Terminal Olefins," *Journal of the American Chemical Society* 131 (2009): 16630–16631; and S. J. Meek, R. V. O'Brien, J. Llaveria, R. R. Schrock, and A. H. Hoveyda, "Catalytic *Z*-Selective Olefin Cross-Metathesis for Natural Product Synthesis," *Nature* 471 (2011): 461–466.

[13] K. Endo and R. H. Grubbs, "Chelated Ruthenium Catalysts for *Z*-Selective Olefin Metathesis," *Journal of the American Chemical Society* 133 (2011): 8525–8527; and B. K. Keitz, K. Endo, M. B. Herbert, and R. H. Grubbs, "*Z*-Selective Homodimerization of Terminal Olefins with a Ruthenium Metathesis Catalyst," *Journal of the American Chemical Society* 133 (2011): 9686–9688.

[14] Even though we evaluate only the reasonable possibilities, we use a lot of computer time. Last year we used about ten million hours of fast computer time, equivalent to one thousand years on one fast computer.

[15] P. Liu, X. Xu, X. Dong, B. K. Keitz, M. B. Herbert, R. H. Grubbs, and K. N. Houk, "*Z*-Selectivity in Olefin Metathesis with Chelated Ru Catalysts: Computational Studies of Mechanism and Selectivity," *Journal of the American Chemical Society* 134 (2012): 1464–1467.

[16] L. E. Rosebrugh, M. B. Herbert, V. M. Marx, B. K. Keitz, and R. H. Grubbs, "Highly Active Ruthenium Metathesis Catalysts Exhibiting Unprecedented Activity and *Z*-Selectivity," *Journal of the American Chemical Society* 135 (2013): 1276–1279.

[17] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, "Towards the Computational Design of Solid Catalysts," *Nature Chemistry* 1 (2009): 37–46.

[18] For a review of this procedure, see G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, and K. N. Houk, "Computational Enzyme Design," *Angewandte Chemie International Edition* 52 (2013): 5700–5725.

[19] See http://www.pdb.org/; and H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research* 28 (2000): 235–242.

[20] See D. Rothlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker, "Kemp Elimination Catalysts by Computational Enzyme Design," *Nature* 453 (2008): 190–195; L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker, "De Novo Computational Design of Retro-Aldol Enzymes," *Science* 319 (2008): 1387–1391; and J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St. Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, and D. Baker, "Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction," *Science* 329 (2010): 309–313.

[21] William L. Jorgensen, "The Many Roles of Computation in Drug Discovery," *Science* 303 (2004): 1813–1818.

[22] National Science and Technology Council, "Materials Genome Initiative for Global Competitiveness" (Washington, D.C.: Executive Office of the President of the United States, 2011).

[23] Andreas W. Götz, Mark J. Williamson, Dong Xu, Duncan Poole, Scott Le Grand, and Ross C. Walker, "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born," *The Journal of Chemical Theory and Computation* 8 (2012): 1542–1555.

[24] Satoshi Maeda and Keiji Morokuma, "Communications: A Systematic Method for Locating Transition Structures of A+B → X Type Reactions," *The Journal of Chemical Physics* 132 (24) (2010): 241102.

From the Atom to the Universe: Recent Astronomical Discoveries

Jeremiah P. Ostriker

Astronomy starts at the point to which chemistry has brought us: atoms. The basic stuff of which the planets and stars are made is the same as the terrestrial material discussed and analyzed in the first set of essays in this volume. These are the chemical elements, from hydrogen to uranium. Hydrogen, found with oxygen in our plentiful oceanic water, is by far the most abundant element in the universe; iron is the most common of the heavier elements. All the combinations of atoms in the complex chemical compounds studied by chemists on Earth are also possible components of the objects that we see in the cosmos. Although almost all of the regions that we astronomers study are so hot that the more complicated compounds would be torn apart by the heat, some surprisingly unstable organic molecules, such as cyanopolyynes, have been detected in cold regions of space with very low density of matter. Nevertheless, the astronomical world is simpler than the chemical world of the laboratory or the real biological world.

But the enormous spatial and temporal extent of the cosmos allows us – and in fact forces us – to ask questions that would seem offbeat to a chemist. Where do the chemical elements come from? Precisely how, where, and when were they made? Do the abundances of the elements change with time? Does alternative "matter" that is not made of the ordinary chemical elements exist and exert gravity in the universe? We in the trades of astronomy and astrophysics *must* ask ourselves these questions – and they are only the beginning.

JEREMIAH P. OSTRIKER, a Fellow of the American Academy since 1975, is the Charles A. Young Professor Emeritus of Astrophysics at Princeton University and Professor of Astronomy at Columbia University. His research interests concern dark matter and dark energy, galaxy formation, and quasars. His publications include *Heart of Darkness: Unraveling the Mysteries of the Invisible Universe* (with Simon Mitton, 2013) and the volumes *Formation of Structure in the Universe* (edited with Avishai Dekel, 1990) and *Unsolved Problems in Astrophysics* (edited with John Bahcall, 1997). He was the recipient of the U.S. National Medal of Science in 2000.

doi:10.1162/DAED_a_00306

In our first essay, "Reconstructing the Cosmic Evolution of the Chemical Elements," Anna Frebel asks precisely the first of these questions. A discoverer of some of the oldest and most metal-poor stars, she tells us how we have found out where and when the nuclear cooking of the elements occurred and precisely which cosmic explosions spewed out which of our familiar elements, from the sodium in salt to the gold in our jewelry. She also introduces some of the remaining mysteries of element creation in the early universe: what do we *not* know?

Let us move beyond standard units of ordinary matter to some larger objects in the universe. The Earth, our beloved planet, is but a grain of sand on the scale of the cosmos; however, it is a respectable body in our solar system of eight normal planets. For literally thousands of years the brighter of these "wanderers" (the meaning of the Greek word "πλανεται," or "planētai") puzzled our ancestors, who believed that their motions among the fixed stars foretold events on earth. The revolutionary discoveries of Tycho, Kepler, Galileo, and Newton identified our position as the third orbiter of the sun, following, along with our companion planets, precisely the paths predicted by Newton's laws of gravity and motion. But are there other planets outside of our solar system? Are the stars that we see in the night sky orbited by their own planets? When I was a graduate student, this was a subject of speculation; there was no factual knowledge. But in the last decades several independent techniques have been developed that tell us without equivocation that extrasolar planets are common! In fact, most stars probably have planets orbiting them. In "Exoplanets, 2003–2013," Gáspár Bakos lays out the dramatic tale of how we have recently found a startling variety of new planets: fat and thin, in round and in elliptical orbits, massive and lightweight. Amazingly, much of the discovery work has been done with relatively small telescopes, using the astronomical equivalent of crowdsourcing (a movement that Bakos has helped lead). The search may soon reach the point where we will be able to use large telescopes to analyze the spectra from the most interesting newly discovered planets to tell if any of them have atmospheres like ours on Earth; and then the question of "life on other worlds" will become the province of science rather than science fiction.

Next up the scale from planets is, of course, the domain of the stars – those fixed points of light in the sky that the ancients cataloged and arranged into houses or constellations. By the eighteenth century it was known that they were not, in fact, fixed, but varied in brightness and moved (albeit slowly) on their own paths across the sky, their current positions by then being significantly different from those recorded by the classical astronomers. By the twentieth century, the enigmatic "nebulae," including the common spiral nebula, were found to be simply giant assemblages of stars and galaxies – "billions and billions of stars," as incanted by the wonderful popularizer of science of the last century, Carl Sagan. We live in such a galaxy – the Milky Way – and our neighbor Andromeda, seen in the northern hemisphere in the winter sky, is another fine example of a typical spiral galaxy. The greatest classical astronomer, Hipparchus, constructed a catalog of fixed stars that had fewer than one thousand entries. By the early twentieth century, the standard HD catalog of bright stars (named for the American astronomer Henry Draper) contained over two hundred thousand entries. The Messier catalog of nebulae published in 1775 contained slightly more than one hundred objects; by the end of the nineteenth century, the similar NGC catalog contained nearly ten thousand galaxies and clusters. But now we live in the age of electronic detectors, huge telescopes, giant computers, and enormous databases. A gigantic explosion of information has occurred in our age of "big data," as outlined in the essay by Michael Strauss, "Mapping the Universe: Surveys of the Sky as Discovery Engines in Astronomy." Strauss, a leader of the Sloan Digital Sky Survey (the largest such survey completed to date), notes that catalogs now contain over a billion stars and galaxies; and they are growing – if the reader will forgive the pun – at an astronomical rate! Couple this with the new tools available for querying databases (such as Google) and one can imagine the rate of discovery.

Massive black holes are much heavier (and stranger) than any star. The first solution indicating the possibility of the existence of black holes was obtained shortly after Einstein invented general relativity in 1915, but it was decades before their character was understood and still longer before black holes were found in nature. Given their common name by the visionary physicist John Wheeler in 1967, they can form when massive stars collapse to such a small size that gravity overwhelms any pressure or nuclear forces, crushing the star into a singularity from which nothing, not even light, can escape. But gaseous matter falling into black holes is heated as it is compressed and will copiously emit light before it disappears into the abyss. This makes black holes visible to astronomers, and many have been found in our galaxy in binary star systems, each one with a mass roughly ten times that of the sun. When quasars – enormously luminous objects at the centers of distant galaxies – were discovered, their variability and incredible luminosity immediately led to speculations that they were much more massive black holes. We now know that the centers of most massive galaxies in fact harbor these enigmatic beasts whose individual masses typically range from four million solar masses at the center of our own galaxy to six billion solar masses at the center of the giant elliptical galaxy M87. The processes by which these megamonsters were formed are under intense investigation and are still quite uncertain. But the saga of how we discovered these extraordinary objects and what we now know about them can be told; and we are fortunate to have had one of the discoverers of quasars, Scott Tremaine, author the essay "The Odd Couple: Quasars and Black Holes." We are just learning that there appear to be close relations between galaxies and their resident central massive black holes, but how and why these relations were formed remains a total mystery.

Moving farther up the cosmic scale, we find galaxies, which are typically a thousand times the mass of the black holes that they harbor at their centers. Galaxies are the basic building blocks of the universe. While it is true that they are collections of stars, they also seem to be embedded in massive halos of mysterious "dark matter," the total weighing in typically at roughly a trillion solar masses with most of it in the mysterious dark component. The visible galaxies were taken for granted by Hubble and the early twentieth-century astronomers as being simply "there," but by the 1960s, the realization had spread that they must be evolving with time, and in fact their formation itself was a subject that must be addressed. Luckily, our increasingly powerful telescopes can see farther and farther out and consequently look back to earlier and earlier times due to the finite speed of light. The most distant galaxies that we can find are thus seen as they existed several billion years ago. With major telescopes, we can use the universe as a time machine and directly study the evolution – and perhaps even the formation – of galaxies in the distant past. Pieter van Dokkum has done just that; in his essay, "The Formation and Evolution of Galaxies," he will limn out what we have learned through the use of powerful telescopes and giant computers that simulate the physics of galaxy evolution. The discoveries are piling up at a great rate in the last decade and we now know that for the most massive galaxies, a two-phase evolution seems to occur: when the galaxy first forms, cold streams of gas converge, flowing into deep wells where the cosmic gravity is greatest due to dark matter accumulations, and huge numbers of stars are formed in relatively small regions. Then, at a later epoch, these monster systems eat their neighboring smaller galaxies (massive black holes and all) and grow further – perhaps by a factor of two in mass and four in size without much additional star formation. Cannibalism among the galaxies! The evolution for lower-mass galaxies like the Milky Way and our companion Andromeda is less well understood. Stay tuned.

Finally, climbing up the ladder of distance and time, we approach the largest known scale: the universe itself. During the last two decades, the knowledge of the universe accumulated over the last century has been synthesized into a well-defined global model that fits all cosmic observations to an uncanny degree. While the nature of the two chief ingredients in this model – dark matter and dark energy – remains a mystery to us, the model has passed all tests given to it so far. The existence of dark matter has been confirmed by both gravitational lensing and the growth patterns of various cosmic structures. Similarly, dark energy seems to have produced the amply observed acceleration of the universe (it is expanding faster and faster rather than slower and slower, as had been expected). But the primary tool for refining and precisely testing the new model has been the analysis of the microwave background (Cosmic Background Radiation, or CBR): the relic radiation left over from the Big Bang. The angular fluctuations seen in the sky by ever more sensitive and precise satellites studying the CBR ("COBE," launched in 1989; "WMAP," launched in 2001; and now "Planck," launched in 2009) have essentially banished doubt about the essential correctness of the standard model of cosmology. David Spergel, a leader of the WMAP satellite team, authors the final essay in this volume, "Cosmology Today," which tells the story of how these results came about, what they tell us about the universe, and what (large) puzzles still remain.

These six essays lay out for the interested reader the extraordinary renaissance that astronomy has undergone in the last decades. We are fortunate indeed to have firsthand discoverers' tales of these adventures to entertain and inform us.

Reconstructing the Cosmic Evolution of the Chemical Elements

Anna Frebel

Abstract: The chemical elements are created in nuclear fusion processes in the hot and dense cores of stars. The energy generated through nucleosynthesis allows stars to shine for billions of years. When these stars explode as massive supernovae, the newly made elements are expelled, chemically enriching the surrounding regions. Subsequent generations of stars are formed from gas that is slightly more element-enriched than that from which previous stars formed. This chemical evolution can be traced back to its beginning soon after the Big Bang by studying the oldest and most metal-poor stars still observable in the Milky Way today. Through chemical analysis, they provide the only available tool for gaining information about the nature of the short-lived first stars and their supernova explosions more than thirteen billion years ago. These events set in motion the transformation of the pristine universe into a rich cosmos of chemically diverse planets, stars, and galaxies.

ANNA FREBEL is the Silverman (1968) Family Career Development Professor in the Physics Department at the Massachusetts Institute of Technology and a member of the MIT Kavli Institute for Astrophysics and Space Research. Her research interests include stellar archaeology and near-field cosmology. Her work has recently appeared in *Nature*, *The Astronomical Journal*, and *Astronomy and Astrophysics*, among other journals; and in the volume *Planets, Stars, and Stellar Systems* (ed. Gerard Gilmore, 2012).

One beautiful afternoon I went for a run along the river. I was breathing plenty of fresh air, my face was all flushed, and I felt my heart pounding and blood flowing through my body. As air was filling my lungs, I was reminded of Carl Sagan's saying: "We are all made from star stuff." Indeed we are. When quenching my thirst with water, I was consuming hydrogen and oxygen in the form of H_2O. When breathing, I had been taking in air made from nitrogen, oxygen, and tiny traces of other elements such as argon and neon. The red liquid of life owes its color to iron, which is embedded in our hemoglobin. But these elements do not just circulate within our carbon-based bodies: before they became part of humans, each of these atoms was created in a grand cosmic cycle called chemical evolution that took place long before biological evolution led to life on Earth.

Most of the universe's iron, for example, is the end result of a binary star system in which one star acquires enough material from its companion that it reaches a critical mass and erupts in a huge thermonuclear explosion, forging new elements in the pro-

doi:10.1162/DAED_a_00307

cess. On the other hand, the hydrogen atoms that make up water are probably nearly fourteen billion years old and were created as part of the Big Bang. And all the carbon upon which life as we know it is based was synthesized in evolved stars near the end of their lives.

The fact that all elements except hydrogen, helium, and lithium are made in stars and their subsequent explosions has only been known for less than sixty years. A seminal paper from 1957, often referred to as "B^{2}FH" following the initials of the authors, provided the first comprehensive summary of "the synthesis of the elements in stars."[1] This came after decades of work directed at finding the energy source of stars. With the elucidation of how and where the chemical elements are forged in stars came the realization that there is a chemical evolution in the universe, causing a net increase in the amount of elements over time. Most important, this model provided observationally testable support for the Big Bang theory and the theory of a time-dependent chemical evolution of the universe.

While the nature of chemical evolution of galaxies is now well established, many details of the complex circle of nucleosynthesis in stars, later chemical enrichment of interstellar gas, and subsequent star formation remain poorly understood and thus continue to be subject to ongoing research. Many questions center around what the exact abundance yields of individual supernova explosions may be, as well as how the nature of the exploding stars themselves and the astrophysical environment influences nucleosynthesis and the production of the elements throughout the periodic table. Because old stars that formed in the early phases of chemical evolution can help with this quest, we will start the tale of the origin of the elements from the very beginning of the universe.

Immediately after the Big Bang 13.8 billion years ago, there was a time without stars and galaxies. The hot gas left over from the Big Bang had to cool enough before the first cosmic objects were able to form. This process took a few hundred million years, but eventually the very first stars lit up the universe. The universe at that time was made from just hydrogen and helium: heavier elements did not exist yet. As a consequence of a variety of gas chemistry and cooling processes that govern star formation, the first stars are thought to have been rather massive. Recent computations suggest these behemoths may have had up to one hundred times the mass of the sun.[2] In comparison, most stars today are low-mass stars with less than one solar mass.

Stars are powered by the nuclear fusion taking place in their cores; it is the energy source that sustains their enormous luminosities. In the first and by far the longest burning phase, hydrogen is fused into helium. At about ten million degrees F, four protons (or hydrogen nuclei) are fused together in a series of nuclear reactions to make a helium nucleus. Subsequent burning stages, which occur only in the last ten percent of stars' lives, result in three helium nuclei ("α-particles") being converted to beryllium, which then captures another particle to become carbon in the so-called triple-α process.

After that, through additional particle captures, carbon nuclei are converted to oxygen; through yet more nucleosynthesis processes, all elements in the periodic table up to iron are built up. The fusion of lighter elements into heavier ones results in a conversion of a small amount of mass into energy. For example, a helium nucleus is 0.7 percent lighter than four individual hydrogen nuclei. It is this mass difference that, as described by $E = mc^2$, fuels the star and sustains its luminosity for long periods of time. However, once the star has created an iron core at its center, nucleosynthesis stops. No more energy can be gained by fusing iron into even heavier nuclei: the star's energy source has ceased for good. As a consequence, the star can no longer maintain equilibrium and begins to collapse due to its own gravity. As a result of the huge pressures, the iron core is converted into an extremely dense neutron star. The collapsing mass of the star bounces off the hard neutron star and leads to a gigantic supernova, leaving the neutron star behind. This was also the fate of the massive very first stars. To sustain their great luminosities, massive stars (those with more than ten solar masses) require large amounts of nuclear energy. Consequently, they burned through the hydrogen and subsequently created heavier-element fuel much more quickly than stars with lower masses, therefore limiting their lifetimes to just a few million years.

During the explosion of a star, all the newly created elements are released into the surrounding gas. The death of the first stars marked an important milestone in the evolution of the universe: it was not pristine anymore, but "polluted" with carbon, oxygen, nitrogen, iron, and other elements. Thus, over time, the universe became more and more enriched in the elements heavier than hydrogen and helium, which are collectively called "metals" by astronomers. In contrast, the very first stars were the only ones that formed from completely metal-free gas. All stars in subsequent generations would then form from gas clouds that contained some metals provided by at least one previous generation of stars exploding as supernovae.

The sudden existence of metals in the early universe following the death of the first stars changed the conditions for subsequent star formation. Gas clouds can cool down more efficiently when metals or dust made from metals are present, leading to the collapse of smaller clouds, and thus the formation of smaller stars. Lower-mass stars like the sun could therefore form for the first time. The first low-mass stars (those with 60 to 80 percent of the mass of the sun) have long lifetimes of fifteen to twenty billion years due to their sparse consumption of the nuclear fuel in their cores. Born soon after the Big Bang as second- or third-generation stars, they are still shining today. Many of these ancient survivors are suspected to be hiding in our Milky Way galaxy and, indeed, astronomers have discovered dozens of them over the past three decades. What makes these extremely rare objects so valuable is that they preserve in their atmospheres information about the chemical composition of their birth cloud, which existed soon after the Big Bang. Hence, studying their chemical composition allows astronomers to reconstruct the early era of their births.

In the earliest stages of the universe's development, massive stars exploding as supernovae dominated the production of iron in the universe. However, this changed after about a billion years. Through the existence of the first lower-mass stars with longer lifetimes, a different pathway for iron production emerged. At the end of their long lives, low-mass stars turn into compact white dwarf remnants. If a star and a white dwarf are in a binary system and enough mass is transferred from the star to the white dwarf, the latter will undergo a thermonuclear explosion. Given the dominance of low-mass stars in the universe today, iron is thus mainly produced by this process rather than by exploding massive stars, as was exclusively the case in the early universe.

After about nine billion years of this chemical evolution, driven by different types of stars at different times, our sun, together with its planets, finally formed. Its birth gas had been enriched by perhaps a thousand generations of stars and supernova explosions. That evolution provided the gas with enough metals to enable the formation of planets – something that may not have been possible much earlier on in the universe. Consequently, when astronomers look for extrasolar planets, they focus their search on stars that are close in age to or younger than the sun.

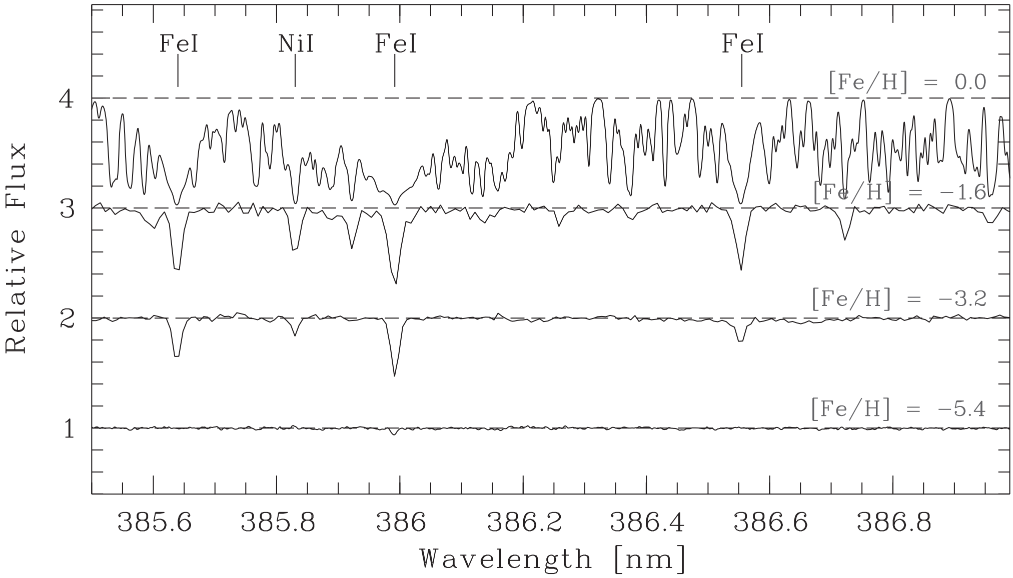
Through spectroscopic observations, astronomers can determine which elements are present in a star's outer layers and what their respective abundances are. Spectroscopy is a technique in which starlight is split up into its components, just as sunlight is split when we see a rainbow. The different elements (hydrogen, helium, and metals) in the star's atmosphere absorb light at very specific colors, or wavelengths. When carrying out high-resolution spectroscopy, the starlight is significantly stretched out over all visible wavelengths to enable the detection of even very weak absorption lines left behind by all the elements in the stars. The existence and strength of absorption lines corresponding to specific elements are measured and analyzed with computer programs that reconstruct the stellar atmosphere. This way, astronomers can calculate how many atoms of a given element are present in the star.

High-resolution spectroscopy, especially for fainter stars, requires the largest telescopes that observe the visible wavelength range. Telescopes like the Magellan-Clay Telescope at Las Campanas Observatory, located in Chile's Atacama desert, are equipped with high-resolution spectrographs. Thanks to its large 6.5-meter-diameter mirror, the Magellan Telescope is capable of collecting enough light from faint stars to enable high-resolution spectroscopic measurements. Chemical analysis then shows how much of each type of metal is present in a star, which indicates the star's formation time. So-called metal-poor stars are assumed to be old because they formed from gas enriched with only a trace amount of heavy elements, created by the first few stellar generations after the Big Bang.[3] In contrast, "metal-rich" stars like the sun must have formed at a much later time when the universe was significantly enriched with metals by many stellar generations.

Our study of early star formation and chemical evolution relies on our ability to measure stars' metallicity, or metal content. The main indicator used to determine stellar metallicity is iron abundance, which, with few exceptions, reflects a star's overall metallicity fairly well. Absorption lines of iron (Fe) can be found throughout stellar spectra, often covering large wavelength ranges from 350 to 900 nanometers, which makes measuring iron abundance relatively straightforward. The iron abundance of a star is given as [Fe/H], which is used in the logarithm of the ratio of iron atoms to hydrogen atoms in comparison to that of the sun. The formal definition reads $[\mathrm{Fe/H}] = \log_{10}(N_{\mathrm{Fe}}/N_{\mathrm{H}})_{*} - \log_{10}(N_{\mathrm{Fe}}/N_{\mathrm{H}})_{\odot}$ with N being the number of Fe and H atoms, respectively, and $*$ and $\odot$ representing the star being evaluated and the sun, respectively. The consequence of this logarithmic definition is that metal-poor stars will have negative [Fe/H] values, as those stars have a lower concentration of Fe atoms than the sun. Stars containing higher concentrations of metals than the sun will show a positive [Fe/H] value.

To illustrate the difference between younger metal-rich and older metal-poor stars, Figure 1 shows spectra of the sun and three metal-poor stars. Their decreasing metallicities are listed. The corresponding number of absorption lines detectable in the spectra decreases with increasing metal-deficiency. In star HE 1327–2326 (bottom spectrum), only very few metal absorption lines are left to observe. Their weakness is such that determining their metal-

Figure 1
Spectral Comparison of the Sun with Three Metal-Poor Stars with Different Metallicities Indicating the Course of Chemical Evolution from the Early Universe (Bottom) to the Sun's Birth ~4.5 Billion Years Ago (Top)



Source: Anna Frebel, "Stellar Archaeology: Exploring the Universe with Metal-Poor Stars," *Astronomische Nachrichten* 331 (2010): 474.

licity requires extremely high-quality data that can only be obtained with large telescopes.[4]

If one wishes to identify the oldest stars, the task is to find stars with the lowest metallicities and thus the earliest formation times. It is those stars that allow astronomers to look back in time and reconstruct the formation and evolution of the chemical elements and the involved nucleosynthesis processes that created them. While very distant galaxies are often used for observational studies of galaxy formation and cosmology, metal-poor stars are the local equivalent of the distant universe and thus the object of "near-field" cosmology. Both approaches to cosmology complement each other in providing detailed information and observational constraints that push us toward understanding the onset of star and galaxy formation in the early universe some thirteen billion years ago. Metal-poor stars are, however, the only tool we have available to learn about the nature of the first stars and their supernova explosions. Our study of these stars therefore provides unique constraints on various theoretical concepts regarding the physical and chemical nature of the early universe.

Past sky surveys for metal-poor stars have shown that these ancient objects can be systematically identified in a three-step process that involves the selection of candidates from the survey data and subsequent follow-up of the best targets with medium- and high-resolution spectroscopy.[5] This technique has identified large numbers of metal-poor stars on the outskirts of the Milky Way, the so-called halo of the galaxy. Work done over the last few decades has shown that stars with low metallicity are much fewer in number compared to more metal-rich stars, reflecting not only the chemical evolution of the universe but also the overwhelming number of stars that have formed since its early stages.

The most metal-deficient stars, in particular, are extremely rare and difficult to find. Only about fifty stars are known to have metallicities of [Fe/H] < –3.5, which corresponds to ~1/3000th of the solar metallicity. Of those, only six have [Fe/H] < –4.0 or < 1/10,000th of the solar value. The current record holder is the star SMSS 0313-6708, with [Fe/H] < –7.0. No iron lines could be detected, so only an upper limit on the iron abundance could be determined, which corresponds to less than ~1/10,000,000th that of the sun. The next most iron-poor star, HE 1327–2326, has an iron abundance of [Fe/H] = –5.4 (~1/250,000th of the solar iron abundance). This translates into an actual iron mass of just 1 percent of the iron mass present in the Earth's core. This is a very small amount, considering that the star is approximately 300,000 times more massive than the Earth and about one million times larger in size. It also reveals that in the early universe, iron and other elements were rare commodities.

Thus, the few stars with [Fe/H] < –4.0 have opened a new and unique observational window to the time shortly after the Big Bang when only the very first stars had enriched the universe. They are frequently employed to constrain theoretical studies about the formation of the first stars, element production and chemical evolution, and supernova yields. The elemental-abundance patterns (chemical abundances as a function of atomic number of the respective element) of these stars appear to be highly individual, but in fact can be successfully reproduced by scenarios in which a massive first supernova explosion provided the elements to the gas cloud from which the observed object later formed. In fact, the most metal-poor stars all display the "fingerprint" of one single massive first supernova, which allows astronomers to ascertain the mechanisms and details of the supernova itself and the nature of the long-extinct progenitor star.

With the exception of hydrogen and helium, all elements up to iron are created through nuclear fusion during lifetimes of stars. But these elements (with atomic number Z ≤ 30) make up less than one third of the periodic table. So where do the other elements with higher atomic masses, such as silver (Z = 47) and gold (Z = 79), or more exotic rare earth elements, such as lanthanum (Z = 57) and europium (Z = 63), come from?

The study of metal-poor stars has greatly advanced our understanding of this topic. As we now know from nuclear physics, elements heavier than iron are created not through fusion processes but through neutron-capture by seed nuclei (for example, iron nuclei). In an astrophysical environment that provides a constant flux of neutrons, heavy elements can thus be built up. Such conditions are thought to occur during certain kinds of supernova explosions in which a strong neutron flux develops above the newly formed central neutron star. For example, if iron nuclei are extremely rapidly bombarded with many neutrons before the nuclei β-decay, their nuclei capture more neutrons, creating heavy, neutron-rich, and unstable isotopes. Once these have β-decayed to stability, new and heavier elements remain.[6] Beta decay is a spontaneous decay of one element into another through the conversion of a neutron into a proton accompanied by the emission of an electron and a neutrino. Due to the rapid bombardment, this process is called the *r*-process. About half of all stable isotopes of elements heavier than zinc are made this way.

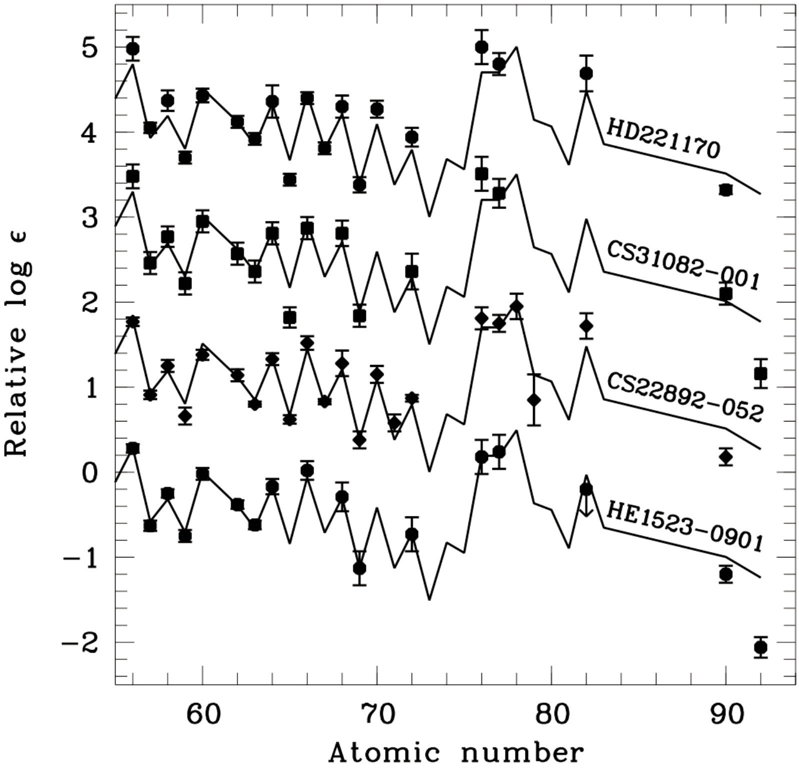
The other half of the isotopes of heavy elements are created in the so-called *s*-process, where a slower neutron bombardment (over a longer timescale than the β-decay process) leads to the successive build-up of heavy elements. This process occurs in the pulsing outer shells of evolved red giant stars with masses of less than eight solar masses and metallicities of [Fe/H] > –3.0 (indicating that the star formed from gas already containing a small amount of iron atoms that could function as seed nuclei). Through stellar winds, these elements are eventually released into the surrounding gas.

In 1995, a low-mass, metal-poor star, CS 22892-052, was discovered to possess a very high abundance of numerous neutron-capture elements (including the rare earth elements) compared to lighter elements such as iron. Indeed, the star has a metallicity of [Fe/H] ~–3.0 (or ~1/1000th of the solar iron abundance), but the neutron-capture material is about forty times more abundant. The various neutron-capture elements detected in this star are likely the result of an *r*-process event that took place prior to the star's birth. When the star formed, it inherited the chemical signature of this particular nucleosynthesis event. For the 4.5-billion-year-young sun, which formed from gas enriched by many generations of stars, it is possible to infer how much of each observed element may have been produced by *r*-process events prior to the sun's formation. The resulting solar *r*-process pattern can be compared to that of other, more metal-poor stars. A comparison between the sun and the metal-poor CS 22892-052, for example, revealed that both stars have the exact same relative pattern of neutron-capture abundances (see Figure 2). It appears that at any time and place in the universe, the *r*-process creates its heavy elements in the exact same ratios, indicating that the *r*-process is a universal process. Since most neutron-capture elements are too heavy to be created and studied in accelerator laboratories on Earth, this has been an important empirical finding based on stellar astronomy.

The elements produced in the *r*-process include thorium and uranium, which are very long-lived radioactive isotopes: ^{232}Th has a half-life of 14 billion years while ^{238}U has a half-life of 4.5 billion years (they are thus decaying very slowly and are near-stable on Earth). Measuring thorium and uranium abundance in metal-poor stars whose birth gas cloud was enriched by only one or few supernova events enables astronomers to carry out cosmo-chronometry: dating the oldest stars with a method analogous to dating archaeological finds through radiocarbon analysis. In the latter technique, the initial ratio of ^{12}C (the typical stable form of carbon) to ^{14}C (an unstable isotope) must be estimated and then compared to the ratio at the time of discovery. Cosmo-chronometry requires astronomers to know the initial amount of the heaviest elements, which were presumably produced together in a massive supernova explosion (obtaining such information is extremely challenging, but detailed calculations of *r*-process nucleosynthesis have yielded estimates). The estimated initial ratios of unstable and stable rare earth elements (such as thorium to europium, uranium to osmium, and thorium to uranium) can then be compared with the currently observed ratios, and the degree of decay of the unstable isotopes thus provides the age of the star.

While thorium is often detectable, uranium poses a great challenge. Only one extremely weak absorption line of uranium is available in the optical spectrum, making its detection difficult, if not impossible,

Figure 2
Elemental Abundances in Four Metal-Poor Stars with a Relative Overabundance of *r*-Process Elements (Various Symbols) Compared with the Scaled Solar *r*-Process Pattern (Solid Line)



All patterns are arbitrarily offset to allow a visual comparison. Note the remarkable agreement of the metal-poor star pattern and that of the sun for elements heavier than barium ($Z \geq 56$). Source: Anna Frebel and John E. Norris, "Metal-Poor Stars and the Chemical Enrichment of the Universe," in *Planets, Stars and Stellar Systems,* vol. 5, ed. Gerard Gilmore (Amsterdam: Springer, 2012), doi:10.1007/978-94-007-5612-0_3.

in most cases. In an ideal scenario, both radioactive elements are detected so that many ratios of the thorium and/or uranium abundance to those of stable rare earth elements can be compared to model predictions for the yields of the *r*-process event. Indeed, several *r*-process metal-poor stars with metallicities of roughly 1/1000th of the solar value were found to be about 14 billion years old. These include HE 1523–0901, which is only the third metal-poor star in which uranium can reliably be detected. Moreover, HE 1523–0901 can be dated with seven different "cosmic clocks"; that is, abundance ratios containing either thorium or uranium and different rare earth elements. The average age obtained through this analysis is 13.2 billion years; this is consistent with the universe's age of 13.8 billion years, which has been deduced from observations of the cosmic background radiation interpreted with the latest cosmological models. Unfortunately, the range of uncertainty with respect to stellar age is often several billion years. Regardless, cosmo-chronometry confirms that HE 1523–0901 and all other metal-poor stars are ancient and formed soon after the Big Bang during the early phases of chemical evolution.

Through individual age measurements, metal-poor *r*-process stars provide an independent lower limit for the age of the universe. This makes them vital probes for near-field cosmology. At the same time, given their rich inventory of very heavy, exotic nuclei, these stars also closely connect astrophysics and nuclear physics by acting as a "cosmic lab" for both fields of study.

Recent searches for metal-poor stars have not only focused on the old stellar halo but also on dwarf satellite galaxies orbiting the Milky Way. The ultra-faint dwarf galaxies – whose total luminosities range from 1,000 to 100,000 solar luminosities, making them the dimmest galaxies known – appear to contain almost exclusively metal-poor stars. These systems ran out of gas for additional star formation billions of year ago. Chemical evolution and star formation ceased as a result, and when we observe these systems, we can only see the leftover low-mass stars that are still shining today. They, too, tell us the story of nucleosynthesis and enrichment in the early stages of the universe.[7] In fact, there are recent indications that these systems are nearly as old as the universe itself: some of them may be among the first galaxies that formed after the Big Bang. Studying these stars thus offers another chance to reconstruct the initial events of element creation within the first stars and their violent explosions, and the subsequent incorporation of this material into next-generation stars. Moreover, the existence of such old satellites may shed light on the existence of metal-poor stars in the halo of the Milky Way. Predating our own galaxy, these halo stars must have come from somewhere; perhaps they originated from dwarf galaxies when analogous systems were gobbled up by the Milky Way during its assembly process.

Topics like these inspire astronomers to collect additional information about the nature and structure of the galaxy. To chemically characterize the galactic halo in detail, including its streams, substructures, and satellites, wide-angle surveys with large volumes are needed. The Australian SkyMapper Telescope is already mapping the Southern sky. It is optimized for stellar work and is delivering new metal-poor star candidates for which high-resolution spectroscopy will be required. The Chinese LAMOST spectroscopic survey is providing numerous metal-poor candidates in the Northern hemisphere. Studying ever-fainter stars further out in the deep halo of the Milky Way and in faraway dwarf galaxies may become a reality with the light-collecting power of the next generation of optical telescopes, including the Giant Magellan Telescope, the Thirty Meter Telescope, and the European Extremely Large Telescope. These telescopes are currently scheduled for completion around 2020. At this point, only the Giant Magellan Telescope is scheduled to be equipped with the high-resolution spectrograph necessary to study metal-poor stars. Further, GAIA, an astrometric space mission led by the European Space Agency (ESA) that was launched in late 2013, will obtain high-precision astrometry for one billion stars in the galaxy, along with the physical parameters and the chemical composition of many of them. Together, these new data will revolutionize our understanding of the origin, evolution, structure, and dynamics of the Milky Way.

All of these new observations will be accompanied by an increased theoretical understanding of the first stars and galaxies, supernova nucleosynthesis, and the mixing of metals into gas clouds in the early universe, as well as cosmic chemical evolution. New generations of sophisticated cosmological simulations of galaxy fomation and evolution will enable a direct investigation of chemical evolution (in a first-galaxy simulation, for example). Being able to trace the metal production and corresponding spatial distributions will allow astronomers to compare the results with abundance measurements of metal-poor stars in the Milky Way's satellite dwarf galaxies. This way, studying nucleosynthesis and the products of chemical evolution will reveal whether any of the ultra-faint dwarf galaxies are surviving first galaxies and whether the metal-poor galactic halo was assembled from early analogs of today's dwarf satellites billions of years ago.

ENDNOTES

[1] E. Margaret Burbidge, Geoffrey R. Burbidge, William A. Fowler, and Fred Hoyle, "Synthesis of the Elements in Stars," *Reviews of Modern Physics* 29 (1957): 547.

[2] Tom Abel, Greg L. Bryan, and Michael L. Norman, "The Formation of the First Star in the Universe," *Science* 295 (2002): 93.

[3] Anna Frebel and John E. Norris, "Metal-Poor Stars and the Chemical Enrichment of the Universe," in *Planets, Stars and Stellar Systems*, vol. 5, ed. Gerard Gilmore (Amsterdam: Springer, 2012), doi:10.1007/978-94-007-5612-0_3.

[4] Anna Frebel, "Stellar Archaeology: Exploring the Universe with Metal-Poor Stars," *Astronomische Nachrichten* 331 (2010): 474.

[5] Timothy C. Beers and Norbert Christlieb, "The Discovery and Analysis of Very Metal-Poor Stars in the Galaxy," *Annual Review of Astronomy and Astrophysics* 43 (2005): 531.

[6] Christopher Sneden, John J. Cowan, and Roberto Gallino, "Neutron-Capture Elements in the Early Galaxy," *Annual Review of Astronomy and Astrophysics* 46 (2008): 241 – 288.

[7] Anna Frebel and Volker Bromm, "Precious Fossils of the Infant Universe," *Physics Today* 65 (4) (2012), doi:10.1063/PT.3.1519.

[8] Anna Frebel and Volker Bromm, "Chemical Signatures of the First Galaxies: Criteria for One-Shot Enrichment," *The Astrophysical Journal* 759 (2012): 12.

Exoplanets, 2003 – 2013

Gáspár Áron Bakos

Abstract: Cosmologists and philosophers had long suspected that our sun was a star, and that just like the sun, other stars were also orbited by planets. These and similar ideas led to Giordano Bruno being burned at the stake by the Roman Inquisition in 1600. It was not until 1989, however, that the first exoplanet – a planet outside the solar system – was discovered. While the rate of subsequent discoveries was slow, most of these were important milestones in the research on extrasolar planets, such as finding planets around a pulsar (a compact remnant of a collapsed star) and finding Jupiter-mass planets circling their stars on extremely short period orbits (in less than a few Earth-days). But the first decade of our millennium witnessed an explosion in the number of discovered exoplanets. To date, there are close to one thousand confirmed and three thousand candidate exoplanets. We now know that a large fraction of stars have planets, and that these planets show an enormous diversity, with masses ranging from that of the moon (1/100 that of Earth, or 0.01$M_\oplus$) to twenty-five times that of Jupiter (25M_J, or approximately 10,000$M_\oplus$); orbital periods from less than a day to many years; orbits from circular to wildly eccentric (ellipses with an "eccentricity" parameter of 0.97, corresponding to an aspect ratio of 1:4); and mean densities from 0.1 g cm^{-3} (1/10 of water) to well over 25 g cm^{-3}. Some of these planets orbit their stars in the same direction as the star spins, some orbit in the opposite direction or pass over the stellar poles. Observations have been immensely useful in constraining theories of planetary astrophysics, including with regard to the formation and evolution of planets. In this essay, I summarize some of the key results.

GÁSPÁR ÁRON BAKOS is an Assistant Professor of Astrophysical Sciences at Princeton University. His research interests include extrasolar planets, instrumentation (with special focus on small telescopes), and all-sky variability. He has served as Principal Investigator of the HATNet and HATSouth extrasolar planet searches, discovering fifty transiting exoplanets so far.

Several processes have been used to discover exoplanets, but the majority have been found by one of the following four observational methods: 1) radial velocity (RV) variations of the host star; 2) brightness variations due to the transit of the planet in front of its host star(s); 3) brightness fluctuations of a background source caused by the gravitational field of the planet (called *microlensing*); and 4) direct imaging of the planet.[1]

The RV method measures the periodic change in the line-of-sight (radial) velocity of the host star, due to the gravitational pull of the planet as it revolves around the star. In other words, the star circles around the center of mass of the star-planet system because of the planet's pull, and we observe the line-of-sight component of this motion. The change in the RV of the star is measured by observing the

doi:10.1162/DAED_a_00308

Doppler shift of the stellar spectrum: the periodic blue-and-red shift of the starlight. Based on the RV signature of the star, the presence and orbital period of the planet can be established, and under certain conditions, a *minimal mass* for the planet is derived. The inclination of the orbit with respect to the line of sight remains unknown; that is, the planet may be orbiting edge-on, or almost face-on. Typical RV variations (for the star) are: approximately 200 ms^{-1} due to a Jupiter-mass object orbiting a solar-type star on a one-day period orbit, 12 ms^{-1} for the same configuration with 5.2-year period orbit (the period of Jupiter itself), and 0.09 ms^{-1} due to an Earth-mass planet orbiting a solar type star on a one-year orbit. Another key parameter measured is the "eccentricity" (ovality) of the orbit. If multiple planets orbit the same star, these parameters can be derived for all the planets.

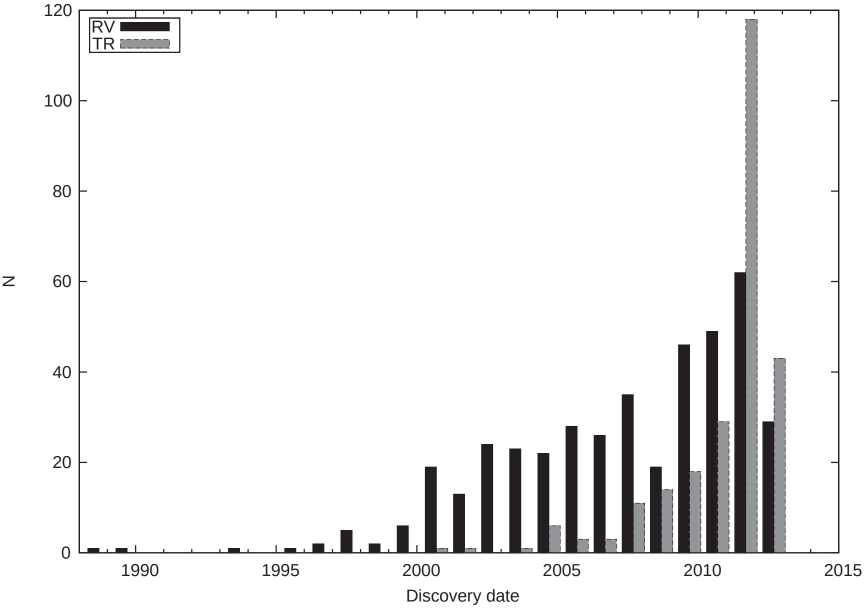
As shown in Figure 1, the number of exoplanet detections has been rising steeply over the past decade. While, in terms of sheer numbers, the RV method was the most successful for much of the past decade, this changed in 2012, when the transit method took over (discussed later). This takeover is even more pronounced if we consider the approximately three thousand planet candidates from the *Kepler* space mission.

While the concept is simple, measuring the RV of a star at the ms^{-1} level has been a challenge, and only a few astronomical facilities have been able to achieve this. Two notable examples, among a dozen facilities, are the High Accuracy Radial velocity Planet Searcher (HARPS) on the European Southern Observatory (ESO) 3.6m diameter telescope, and the HIgh Resolution Echelle Spectrograph (HIRES) on the Keck-I 10m diameter telescope. The precision of instruments has improved significantly over a decade: for example, HARPS reaches 1 ms^{-1} precision for a moderately bright star in a one-minute exposure. For bright stars, the primary limitations on precision include instrument systematics and noise due to stellar activity. Recent record-breaking detections include a planet inducing an RV variation of only a half-meter per second on its host star (called HD 20794) and a possible Earth-mass planet around one of the brightest and closest stars, α Centauri B, causing a similar RV variation roughly equivalent to the speed of a person walking slowly.

Significant advances have been made by way of high-precision spectroscopy using "laser-combs,"[2] which provide a highly accurate and stable calibration source. High-precision *infrared* spectroscopy[3] has also been at the frontier of research, motivated by the enhanced detectability of potentially habitable Earth-mass planets around smaller (and cooler) stars. Plans for future instrumentation on the next generation of large telescopes under development are being shaped by the goal of detecting small planets. (Examples include the ESPRESSO instrument on the 8.2m diameter VLT telescope, CODEX for the future 39m E-ELT telescope, and GCLEF for the future 24m GMT telescope.)

To date, RV searches have targeted a few thousand relatively bright stars, and have discovered around five hundred exoplanets in approximately four hundred planetary systems (some of which are multiplanet systems). This sample is large enough to derive meaningful statistics, as has been done by many authors.[4] Some of the key results include: Gas giant planets with planetary mass $M_p > 50M_\oplus$ (50 Earth masses) on short-period orbits (less than ten days), also known as "hot Jupiters," are intrinsically rare, present in only around 1 percent of all star systems. However, there is an extremely strong bias favoring their discovery, which explains why the first RV detection of a Jupiter-mass planet (around the star 51 Peg) was that of a hot

Figure 1
Number of RV (black) and Transit (gray) Detections as a Function of Year



This plot does not show planet candidates from the Kepler mission. Source: Data from exoplanets.org.

Jupiter,[5] and why most ground-based transit surveys have discovered only hot Jupiters. In contrast, "light" planets more like Earth ($M_p < 30M_\oplus$) are abundant, with a very sharp increase in the occurrence rate as planetary mass decreases. The occurrence rate of giant planets increases with the metal content (fraction of elements heavier than helium; also called "metallicity") and mass of the host star. This, however, is not true for light planets (Neptunes, super-Earths, and smaller), which have a much weaker dependence on metallicity. For giant planets, we observe a bimodal distribution in the period, with a small "pile up" at $P \approx$ three days (hot Jupiters) followed by a "period valley" and steep increase in the occurrence rate for $P \geq$ one hundred days. Light planets, however, tend to have short-period orbits, the most typical period being forty days. While giant planets exhibit a wide distribution of eccentricities (even reaching $e \approx 0.97$), small planets exhibit more circular orbits. Hot Jupiters are "lonely," with either no detectable companion or a companion on a very wide orbit. In contrast, small planets are often in multiplanetary systems[6] like our own solar system. These observational *facts* are extremely important for constraining the various planet formation and evolution theories.

Some interesting numbers on the occurrence rate of exoplanets, as based on RV searches, are as follows: three-quarters of dwarf (solar-like) stars have a planet with a period less than ten years, and one-quarter of dwarf stars have a 0.5 to $2M_\oplus$ Earth-mass planet with a period less than fifty days.[7] It may turn out, when all periods and masses are considered, that essentially all stars have planets.

Below I list some notable exoplanetary systems that have been discovered by the RV method. Note that the nomenclature of exoplanets is such that most exoplanets carry the name of the host star together with a suffix for the planet ("b" for the first, "c" for the second, and so on). For example, the first planet discovered around the star 55 Cnc is called 55 Cnc b, the second and third planets are 55 Cnc c and 55 Cnc d, respectively.

- HD 80606 b is a massive planet on an extremely eccentric and long-period orbit, and was later found to transit its host star.
- A Neptune-mass planet orbiting a nearby M dwarf, GJ 436, was later found to transit its host star.
- Four planets were discovered around the red dwarf star Gl 876,[8] with three of them in the so-called Laplace-resonance, where the ratio of the orbital periods is 1:2:4. This configuration is strikingly similar to the inner three Galilean moons of our Jupiter.
- A system of three to six planets was found around Gl 581 (some of which are disputed), including one or more super-Earths (planets more massive than Earth, but less than about ten times the mass of Earth) close to or inside the habitable zone.
- 55 Cnc is a star visible to the naked-eye with five planets. The innermost planet (55 Cnc e) is a super-Earth on a very short-period orbit of only 0.73 days, and was later found to transit the star.
- HD 10180 has a planetary system of seven (or even more) planets, the largest number yet for exoplanetary systems!
- An Earth-mass planet was discovered, inducing only 0.5 ms^{-1} variation on α Centauri B, one of the brightest and closest stars in the sky.[9] (Note: this discovery is still debated.)

High-precision RV measurements will continue to improve in the next decade, and should reach a precision level of just a few centimeters per second. A serious limitation on the method will be stellar noise, due to either oscillations of the star or spots and other surface irregularities.

By chance alignment, our line of sight may lie in the orbital plane of an exoplanet. In this case, the planet periodically transits across the face of the star as seen from Earth. During transits, the star's light is dimmed by a fraction that is proportional to the ratio of the projected area of the planet to that of the star: that is, $(R_p/R_\star)^2$, where R_p and $R_\star$ are the planetary and stellar radii, respectively. As viewed from such a vantage point, Jupiter transiting the sun would cause a 1 percent dip in the *light curve* (the light of the star as a function of time) lasting roughly thirty hours, and Earth would cause a 0.01 percent dip for about thirteen hours.

Careful analysis of the light curve during the transit, together with RV observations of the system, yield the following parameters for the system: period, planet-to-star radius ratio, inclination (angle or orbital plane with respect to the sky plane), and semi-major axis of the planet (half of the longest diameter of the elliptical orbit, in units of the stellar radius). Further, using Kepler's third law, the following fundamental physical parameters are determined: the surface gravity of the planet and the mean density of the star. When coupled with spectroscopic observations and stellar models, the mean stellar density is, in turn, an important constraint on the mass and radius of the star. In other words, the transiting planet helps us determine the properties of its host star. Knowing the host star, then, is essential for determining the parameters of the planet.

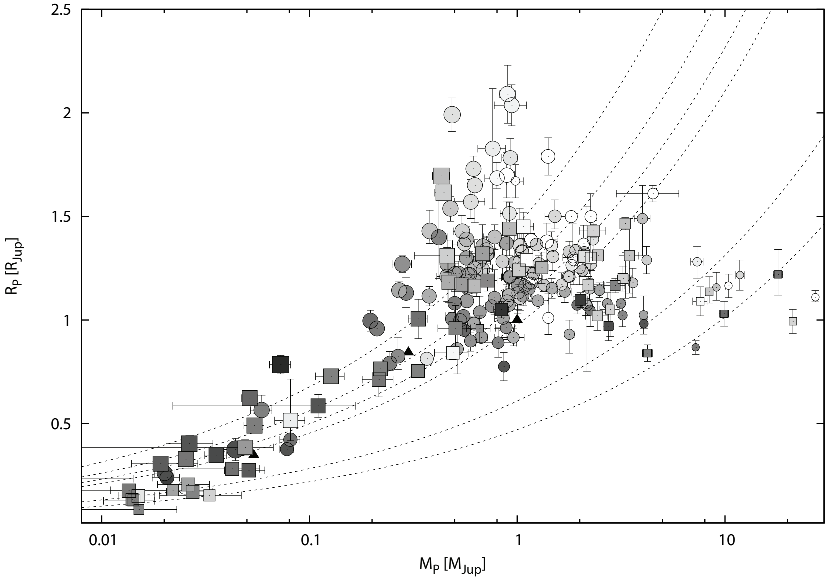
For transiting exoplanets (TEPs), if the stellar radius is known (which is typically the case), the planetary radius is also determined. If RV measurements of the host star are available, and if the stellar mass is known (which is also common), then the mass of the TEP is also derived without any ambiguity (not only a lower limit, as for pure RV detections).

Astronomers recognized long ago that TEPs provide us with an unparalleled opportunity for understanding their physical properties, as one can unambiguously determine their masses, radii, mean densities, and surface gravities, among other measurements. The chance of detecting transit of a planet around a random star, however, was initially thought to be very slim, because 1) giant planets were thought to be rare; 2) the chance that we would fall in its orbital plane is tiny (for example, 1 in 1,000 for our Jupiter, as seen from a vantage point); 3) the fraction of time spent transiting is small (for example, 0.0003 for our Jupiter); and 4) the transit signature (diminution of starlight) is small (< 1 percent). Thus, it is no accident that the first TEP detection, in 2000,[10] was that of a giant planet (HD 209458 b), previously known from RV measurements to orbit a bright star on a very short-period orbit. Only a couple of years later, in 2003, the OGLE project[11] detected the first TEPs *without* prior RV measurements.[12] The enormous scientific value of transiting planets triggered a gold rush for TEPs, and a small armada of projects were undertaken. These employed primarily wide-field instruments observing tens of thousands of stars per exposure and monitoring stellar fields every clear night. Certain challenges were realized, such as the need for 1) robust automation of telescopes and data processing; 2) high-precision photometry (stellar flux measurements) over a wide field in the sky; and 3) extensive resources for following up on the planet candidates to deal with the large number of astrophysical scenarios that mimic planetary transits. Nevertheless, as TEPs became a hot topic in contemporary astrophysics, their study provided a unique opportunity for small telescopes to participate in cutting-edge science. Examples of such surveys include TrES, XO, HATNet, WASP, and HATSouth. Together, these projects have revealed approximately 150 TEPs, which are among the best characterized exoplanets. In fact, the majority (70 percent) of exoplanets with masses and radii measured to better than 10 percent accuracy were discovered by wide-field ground-based surveys.

Figure 2 plots the mass-radius diagram for around two hundred TEPs with well-determined physical parameters. Broadly speaking, the radius increases with increasing mass, but there are a number of interesting features in this figure. One is that – as theoretical physics would predict – the radii of massive planets are relatively small; that is, the planets are very dense. Also, many transiting gas giant planets on short-period orbits were found to have much larger radii than that expected for an old and cool pure hydrogen/helium body (see the top portion of Figure 2). One example is HAT-P-32 b, with mass equal to but a radius twice that of Jupiter's – meaning the planet is 1/8 Jupiter's mean density! There has been no shortage of theories to explain this "inflation" of planets. Ground-based surveys yielded a big enough sample to show that the inflation of radii (relative to those predicted) was connected to the heating by in-falling stellar flux, and perhaps by the metal content of the star. At present there is still no clear understanding of this matter.[13]

During the transit of a planet, starlight passes through the atmosphere of the planet, and a fraction of the light is absorbed, depending on the properties of the atmosphere (chemical composition, scale, height). By comparing the stellar spectrum in and out of transit, one can infer these

Figure 2
The Mass-Radius Diagram of Transiting Extrasolar Planets



Circles indicate ground-based discoveries, while squares show space-based discoveries. Point-size scales with the inverse of the planetary surface gravity (higher gravity = smaller points). The gray tone scales with equilibrium temperature, with the very light symbols indicating temperatures in excess of 2000 K. Dashed lines indicate iso-density lines corresponding to 0.4, 0.7, 1.0, 1.33, 5.5, and 11.9 g cm^{-3}, respectively.

properties of the planetary atmosphere. The first such measurement was the detection of sodium in the atmosphere of the planet HD 209458 b using the STIS instrument on the Hubble Space Telescope (HST). Since then, this planet has been subject to intensive studies, detecting an extended hydrogen exosphere, for example, and water absorption in the near infrared[14] via transmission spectroscopy. Another well-studied system is HD 189733 b, a gas giant on a short period orbit around a nearby star (which star is somewhat cooler than the sun). Here sodium was detected from the ground using the Hobby Eberly Telescope, and a featureless spectrum in the visible (HST/ACS) suggested haze in the atmosphere. Can you imagine the excitement if we were to detect molecular oxygen in the atmosphere of one of the planets? It would suggest that there may be life on the planet, since oxygen is very reactive and does not persist absent living organisms. (Note, however, that this is a potential "biosignature," and not definite evidence.[15])

During the occultation of a planet – that is, when the planet moves *behind* the star – the combined light from the star and planet drops (as seen from Earth). In the *infrared*, this drop is primarily due to the planet's *thermal radiation* being eclipsed by the star. By measuring the depth of the occultation, one can measure the so-called brightness temperature on the "dayside" of the planet. Such measurements have been performed for more than thirty star systems. In the case of the aforementioned HD 189733 b,[16] the "phase-curve" (total observed brightness as the planet moves on its orbit) was observed from before primary transit to after occultation, and night- and dayside brightness was found to reach temperatures of approximately 970 K and 1200 K, respectively. This is a relatively small difference, especially because we believe that this planet is tidally locked: the planet shows the same side to the star (just as our moon shows the same face to Earth at all times), which therefore receives an enormous radiant flux. A plausible solution is that the heat is efficiently redistributed to the nightside via circulation of winds. Even more amazing is that we now have a thermal map of this exoplanet showing that the warmest spot on the planet is 16° off (to the "east") from the point directly facing the star.

Phase-curves were automatically acquired by the *Kepler* space mission for thousands of transiting planet candidates, as it performed nonstop observations in the visible band-pass. Occultations in the *visible* light (as compared to the infrared, discussed above) are primarily due to reflected light occulted by the star, and provide a handle on the reflectivity of extrasolar planets. For the planet TrES-2 b, *Kepler*'s measurements established a stunningly low reflectivity; this planet is "pitch dark," reflecting only 2.5 percent of the infalling light.[17]

Planetary spectra were also investigated close to their occultation. By comparing the spectrum of the system during and outside occultation, one can infer the emission spectrum of the planet. Molecules such as CH_4, H_2O, and CO_2 were detected for a number of exoplanets.

Observing transmission spectra or other properties of exoplanets can be extremely challenging from the ground, due to systematic noise introduced by our own atmosphere. Nevertheless, it has been a trend in recent years to use ground-based instrumentation, such as OSIRIS on the Gran Telescopio Canarias, and multi-object spectrographs with wide slits, such as MMIRS on the Magellan Telescope. In general, astronomers made excellent use of instruments that were not originally designed to carry out such high-precision measurements, occasionally stretching beyond the capabilities of these tools. Results have sometimes been questioned by competing teams, and were shown to be very sensitive to the data analysis procedures.[18]

Gáspár Áron Bakos

As hinted at earlier, another quantity that can be measured for TEPs is the angle between the orbital plane of the planet and the equator of the star, as projected onto the plane of the sky. This angle is revealed through an anomaly in the RV measurements of the star made during transit. This "Rossiter-McLaughlin" anomaly was predicted and observed for eclipsing binary stars almost a century ago, but was only applied to TEPs in the past decade.[19] Initial measurements suggested that all TEP orbits are well aligned ("prograde") with the equator of their host stars, and this almost led to a pause in the investigation of further systems. Then, highly tilted (called *high obliquity*) and even retrograde planets (circling the star in the "opposite" direction) were discovered, such as HAT-P-7 b.[20] This hot Jupiter, on a 2.2-day retrograde orbit, clearly violated the prevailing theory of giant planet formation, which argued that planets form far from the stellar heart of the system, where ices condense from the rotating protoplanetary disc and then slowly migrate inward (keeping at least the direction of their angular momentum). Astronomer Joshua Winn and colleagues concluded that hot stars with hot Jupiters have high obliquities, with the dividing line between well-aligned and misaligned systems being somewhere at stellar effective temperature higher than the sun.[21] Recently, physicist Simon Albrecht and colleagues have claimed that the star-planet obliquities for close-in hot Jupiters were initially random, and aligned systems are those where tidal interactions between the star and the planet are expected to be strong (see Figure 3).[22]

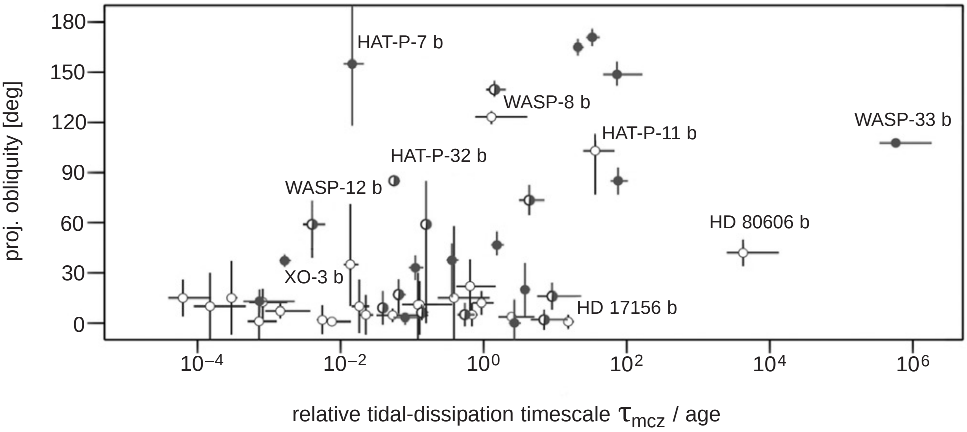
Space-based transit searches are qualitatively different from ground-based searches in that they have achieved much higher photometric precision than ground-based surveys, using almost uninterrupted observations to observe fainter stars. One key player was the CoRoT satellite. Perhaps one of the top scientific results of CoRoT was the discovery of CoRoT-7 b,[23] a transiting super-Earth with $1.6R_\oplus$ radius on a twenty-hour-period orbit, and with a mass $\leq 8M_\oplus$.

The other key player in exoplanet discovery is the *Kepler* space mission, which is fully dedicated to TEP detection. It was designed to have the capability of detecting an Earth-sized planet transiting a sun-like star. *Kepler* has been transformative to the field. Launched in 2009, it has been continuously monitoring a selected area in the sky, roughly the area of the Big Dipper, yielding exquisite photometry for some 150,000 stars. Most stars are observed at a thirty-minute cadence, and the per-point precision of the stellar fluxes reaches thirty parts per million! (*Kepler* failed mechanically in May 2013, and a new mission plan, "K2" or "Second Light," has been adopted to make use of *Kepler*'s remaining capabilities.) To date, *Kepler* has found some three thousand planetary *candidates*. As astrophysicist Timothy Morton and astronomer John Asher Johnson have shown, based on statistical arguments, at least 90 percent of these should be real planets, even though the classical confirmation is not available for the majority.[24]

Notably, the *Kepler* space mission found that small (radius) planets, reaching down to Earth-size, are extremely frequent. Using the *Kepler* data, Andrew Howard and colleagues found that for orbital periods shorter than fifty days, the distribution of planet radii scales with the inverse square of the planetary radius.[25] Approximately 13 percent of stars have small ($2 – 4R_\oplus$) planets with relatively short (< fifty days) periods. They also found that the occurrence of small planets in the *Kepler* field increases for cool stars (< 4000 K) by a factor of seven when compared to hot stars (> 6600 K). It also appears that while smaller planets are more frequent, this occurrence rate plateaus at about $2R_\oplus$; that is, planets smaller than two Earth-radii are still not more frequent. Another important finding of *Kepler* was that multiple planetary systems are intrinsically frequent: at least 27 percent of planets are in multi-transiting systems, and 15 percent of stars with TEPs host more than one planet.[26] Multiple systems consist primarily of small planets; hot Saturns and Jupiters are "lonely," with no nearby companions. A shortlist of some truly amazing planetary systems found by *Kepler* is given below.

- Kepler-11: six TEPs in a densely packed configuration, five with orbital periods between ten and forty-seven days.
- Kepler-36: a pair of planets with orbital distances differing by only 10 percent and with densities differing by a factor of eight.
- Kepler-47, the first circumbinary planetary *system*: two super-Earths orbiting at $P \approx 50$ days and $P \approx 300$ days around a pair of stars that are eclipsing each other about every seven days.
- Kepler-62: a five-planet system with planets of 1.4 and $1.6R_\oplus$ radii orbiting in the habitable zone (that is, they *may* host life).[27]
- KOI-872 b, c: the detection and characterization of a nontransiting planet (KOI-872 c) by variations (due to gravitational interaction) induced on the transit times of the transiting planet (KOI-872 b).

Figure 3
Obliquity of Transiting Exoplanets with Respect to the Spin Axis of Their Host Stars



The vertical axis shows the obliquity of the planets with respect to the stellar spin axis, projected on the sky plane. Zero degrees means a perfectly aligned prograde orbit, and 180 degrees means retrograde orbit. The horizontal axis shows the estimated timescale required for aligning a planet with the stellar spin axis. Stars with temperatures higher than 6250 K are shown with filled symbols. Open symbols show stars with temperatures lower than 6250 K. Source: Simon Albrecht et al., "Obliquities of Hot Jupiter Host Stars: Evidence for Tidal Interactions and Primordial Misalignments," *The Astrophysical Journal* 757 (2012): 18, arXiv:1206.6105; used here with permission of the American Astronomical Society.

- KOI-142: similar case to KOI-872, but also using variations in the transit duration (for the first time) to determine the system parameters.

The next decade will be extremely promising in terms of scientific progress based on transiting planets. The U.S.-based TESS mission, with a proposed launch sometime in 2017, will scan the entire sky to detect thousands of transiting planets around the brightest stars near Earth. Among these will be potentially habitable super-Earths amenable to follow-up observations. The James Webb Space Telescope (JWST) and the European EChO space mission will observe atmospheres of TEPs through transmission and occultation spectroscopy at unprecedented precision. There is hope that by the next decade we will actually detect bio-signatures in the atmospheres of remote worlds.

The past decade saw real breakthroughs in the direct imaging of exoplanets. The task is extremely challenging because stars are bright while planets are faint and appear close to the stars; thus, capturing and separating the light emanating from the planet requires very-high-contrast and high-spatial-resolution imaging. For these reasons, state-of-the-art instrumentation has been developed and used on the largest telescopes, employing infrared imaging, adaptive optics, coronography, and novel observing and data analysis techniques. The targets typically are young (a few million years old) stars, because these may have young planetary systems, in which the planets still emit excess (infrared) light due to their primordial heat.

One spectacular success was direct imaging of the planets around the young star HR 8799.[28] Not only were the planets clearly visible, but their face-on orbital motion around the star was also apparent on images taken a year apart. Another important discovery was the planet orbiting the very bright and nearby star, β Pictoris.[29] This star has a disk of debris (dust and rocks) that is almost edge-on. An approximately 8M_J planet around the same star was detected on archival images from 2003, then reobserved in 2009 on the other side of the star! Finally, a very recent detection is the 13M_J mass single gas giant around the star "κ And," detected with the Subaru/HiCIAO instrument as part of the SEEDS survey.[30]

Another breakthrough was the ability to take *spectra* of the directly imaged planets. This was done for all four planets around HR 8799 by Project 1640 on the venerable Palomar 5m telescope.[31] The low-resolution spectra show an unexpected diversity among the four planets, with hints of CH_4, NH_3, CO_2, and other molecules. Massive current efforts (NICI, SEEDS) and future projects (GPI, SPHERE) will certainly yield many more directly imaged planets with spectra.

An object (star or planet), by virtue of its finite mass, will perturb the light from a background source that falls along the line of sight, creating multiple images and magnifying the source. As the object ("lens") moves with respect to the background source, the magnification changes in time, resulting in the brightening of the background source. The brightening, as the function of time (the light curve), has a characteristic "bell-shape." This effect, predicted by Einstein, has now become a practical astronomical tool. Thousands of such microlensing events[32] are now detected every year by surveys such as OGLE[33] and MOA.[34] Planets around the lensing star cause further perturbation of the light, and appear as anomalies on the microlensing light curve. To date, some thirty or so planets have been discovered via microlensing. An interesting aspect of microlensing is its sensitivity toward low-mass planets on wide orbits, even free-floating planets without a host star. Microlensing is sensitive to classes of planets that currently cannot be found by RV, transit, or direct-imaging searches.

The first microlensing-detected planet was a 2.6 Jupiter-mass object at a 5 AU orbit (astronomical unit, the mean distance between the Earth and the sun: 149.6 million kilometers).[35] Microlensing detected planets more massive than Jupiter around very small stars,[36] which is somewhat surprising given the scarcity of such objects found by RV and transit searches. It is also from microlensing that we learned about the existence of cold super-Earths that orbit their stars beyond the ice-line (distance beyond which it is cold enough for ices to form) at several AU distance. Multiple planetary systems, such as a Jupiter/Saturn analogue, were also detected by microlensing.[37] Recently, microlensing has found signals due to free-floating planets, and has concluded that such planets are twice as common in our galaxy as main-sequence stars.[38]

While progress in the field of exoplanets over the past decade has been spectacular, most of the excitement is yet to come. This includes finding analogues of our solar system, planets similar to our Earth, and moons around exoplanets. We hope to detect biomarkers and – ultimately – signs of intelligent life capable of communication with us, however slow the turnaround time may be.

ENDNOTES

Author's Note: I thank Joel Hartman and Jeremiah Ostriker for careful reading of this manuscript. I apologize to those whose work I could not discuss due to space limitations.

[1] Michael Perryman, *The Exoplanet Handbook* (Cambridge: Cambridge University Press, 2011).

[2] Gaspare Lo Curto et al., "Achieving a Few cm/sec Calibration Repeatability for High Resolution Spectrographs: The Laser Frequency Comb on HARPS," *Proceedings of SPIE* (2012): 8446.

[3] Peter Plavchan et al., "Precision Near-Infrared Radial Velocity Instrumentation and Exoplanet Survey," American Astronomical Society, AAS Meeting No. 221, #109.06, 2013.

[4] For example, Andrew W. Howard et al., "The Occurrence and Mass Distribution of Close-in Super-Earths, Neptunes, and Jupiters," *Science* (330) (2010): 653.

[5] Michel Mayor and Didier Queloz, "A Jupiter-Mass Companion to a Solar-Type Star," *Nature* (378) (1995): 355.

[6] Michel Mayor et al., "The HARPS Search for Southern Extra-Solar Planets XXXIV: Occurrence, Mass Distribution and Orbital Properties of Super-Earths and Neptune-Mass Planets," *Astronomy & Astrophysics* (2011), arXiv:1109.2497.

[7] Andrew W. Howard et al., "Planet Occurrence within 0.25 AU of Solar-Type Stars from Kepler," *The Astrophysical Journal Supplement* (201) (2012): 15.

[8] Eugenio Rivera et al., "The Lick-Carnegie Exoplanet Survey: A Uranus-Mass Fourth Planet for GJ 876 in an Extrasolar Laplace Configuration," *The Astrophysical Journal* (719) (2010): 890.

[9] Xavier Dumusque et al., "An Earth-Mass Planet Orbiting α Centauri B," *Nature* (491) (2012): 207.

[10] David Charbonneau et al., "Detection of Planetary Transits Across a Sun-Like Star," *The Astrophysical Journal* (529) (2000): L45.

[11] Andrzej Udalski et al., "The Optical Gravitational Lensing Experiment: Search for Planetary and Low-Luminosity Object Transits in the Galactic Disk. Results of 2001 Campaign," *Acta Astronautica* (52) (2002): 1.

[12] M. Konacki et al., "An Extrasolar Planet that Transits the Disk of Its Parent Star," *Nature* (421) (2003): 507.

[13] For a review, see David S. Spiegel and Adam Burrows, "Thermal Processes Governing Hot-Jupiter Radii," *The Astrophysical Journal* (2013), arXiv:1303.0293.

[14] D. Deming et al., "Infrared Transmission Spectroscopy of the Exoplanets HD 209458 b and XO-1 b Using the Wide Field Camera-3 on the Hubble Space Telescope," *The Astrophysical Journal* (774) (2013): 95, arXiv:1302.1141.

[15] Hanno Rein, Yuka Fujii, and David S. Spiegel, "Some Inconvenient Truths about Biosignatures Involving Two Chemical Species on Earth-Like Exoplanets," *Proceedings of the National Academy of Sciences* 111 (19) (2014): 6871 – 6875.

[16] Heather A. Knutson et al., "A Map of the Day-Night Contrast of the Extrasolar Planet HD 189733 b," *Nature* (447) (2007): 183.

[17] David M. Kipping and David S. Spiegel, "Detection of Visible Light from the Darkest World," *Monthly Notices of the Royal Astronomical Society* (417) (2011): L88.

[18] For example, see Deming et al., "Infrared Transmission Spectroscopy of the Exoplanets HD 209458 b and XO-1 b Using the Wide Field Camera-3 on the Hubble Space Telescope."

[19] D. Queloz et al., "Detection of a Spectroscopic Transit by the Planet Orbiting the Star HD 209458," *Astronomy and Astrophysics* (359) (2000): L13; and Joshua N. Winn et al., "Measurement of Spin-Orbit Alignment in an Extrasolar Planetary System," *The Astrophysical Journal* (631) (2005): 1215.

[20] Joshua N. Winn et al., "HAT-P-7: A Retrograde or Polar Orbit, and a Third Body," *The Astrophysical Journal* (703) (2009): L99.

[21] Joshua N. Winn et al., "Hot Stars with Hot Jupiters Have High Obliquities," *The Astrophysical Journal* (718) (2010): L145.

[22] Simon Albrecht et al., "Obliquities of Hot Jupiter Host Stars: Evidence for Tidal Interactions and Primordial Misalignments," *The Astrophysical Journal* 757 (2012): 18, arXiv:1206.6105.

[23] D. Queloz et al., "The CoRoT-7 Planetary System: Two Orbiting Super-Earths," *Astronomy and Astrophysics* (506) (2009): 303.

[24] Timothy D. Morton and John Asher Johnson, "On the Low False Positive Probabilities of *Kepler* Planet Candidates," *The Astrophysical Journal* (738) (2011): 170.

[25] Howard et al., "Planet Occurrence within 0.25 AU of Solar-Type Stars from *Kepler*."

[26] Ibid.

[27] William J. Borucki et al., "Kepler-62: A Five-Planet System with Planets of 1.4 and 1.6 Earth Radii in the Habitable Zone," *Science* (340) (2013): 587, arXiv:1304.7387.

[28] Christian Marois et al., "Direct Imaging of Multiple Planets Orbiting the Star HR 8799," *Science* (322) (2008): 1348; and Christian Marois et al., "Images of a Fourth Planet Orbiting HR 8799," *Nature* (468) (2010): 1080.

[29] A.-M. Lagrange et al., "A Probable Giant Planet Imaged in the Beta Pictoris Disk: VLT/NaCo Deep L'-band Imaging," *Astronomy & Astrophysics* (493) (2009): L21.

[30] J. Carson et al., "Direct Imaging Discovery of a 'Super-Jupiter' around the Late B-Type Star κ," *The Astrophysical Journal* (763) (2013): L32.

[31] B. R. Oppenheimer et al., "Reconnaissance of the HR 8799 Exosolar System. I. Near-Infrared Spectroscopy," *The Astrophysical Journal* (768) (2013): 24.

[32] B. Paczyński, "Gravitational Microlensing by the Galactic Halo," *The Astrophysical Journal* (304) (1986): 1.

[33] Andrzej Udalski, " The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey," *Acta Astronautica* (53) (2003): 291.

[34] T. Sako et al., "MOA-cam3: A Wide-Field Mosaic CCD Camera for a Gravitational Microlensing Survey in New Zealand," *Experimental Astronomy* (22) (2008): 51.

[35] I. A. Bond et al., "OGLE 2003-BLG-235/MOA 2003-BLG-53: A Planetary Microlensing Event," *The Astrophysical Journal* (606) (2004): L155 – L158.

[36] Andrzej Udalski et al., "A Jovian-Mass Planet in Microlensing Event OGLE-2005-BLG-071," *The Astrophysical Journal* (628) (2005): L109.

[37] B. S. Gaudi et al., "Discovery of a Jupiter/Saturn Analog with Gravitational Microlensing," *Science* (319) (2008): 927.

[38] T. Sumi et al., "Unbound or Distant Planetary Mass Population Detected by Gravitational Microlensing," *Nature* (473) (2011): 349.

Mapping the Universe: Surveys of the Sky as Discovery Engines in Astronomy

Michael A. Strauss

Abstract: Astronomers can map the sky in many ways: observing in different regions of the electromagnetic spectrum, obtaining spectra of stars and galaxies to determine their physical properties and distances, and repeatedly observing to measure the variability, explosions, and motions of celestial objects. In this review I describe recent surveys of the sky astronomers have carried out, focusing on those in the visible part of the spectrum. I describe in detail the Sloan Digital Sky Survey, an ongoing imaging and spectroscopic survey of over one quarter of the celestial sphere. I also discuss some of the major surveys planned for the next decade, using telescopes both on the ground and in space.

MICHAEL A. STRAUSS is Professor of Astrophysical Sciences and Associate Chair of the Department of Astrophysical Sciences at Princeton University. His research concerns all aspects of extragalactic astronomy and observational cosmology. He has published over two hundred refereed papers on subjects ranging from the large-scale distribution of galaxies to the discovery of the most distant quasars known.

Astronomy is an observational science. Unlike chemistry or biology, the objects of study in astronomy are far removed, at distances to which we will not have the capability to travel using even the most advanced foreseeable technology. This means that we cannot carry out experiments on the stars and galaxies that are the bread and butter of our discipline; all the information we can glean about them is the result of measuring the tiny fraction of the light that they emit that happens to fall on our eyes and our telescopes. We then interpret these data in the context of the laws of physics to draw conclusions about the nature of these distant bodies, allowing us to infer, for example, the conditions in the cores of stars, or the existence of new forms of matter that are unknown from our experience and experiments here on Earth.

The range of phenomena in the universe is vast, and the rate of astronomical discovery today tells us that we are far indeed from a complete understanding of all that the universe has to teach us. This essay describes one of the most productive approaches we have toward astronomical discovery; namely, using our telescopes to map the heavens and create

doi:10.1162/DAED_a_00309

a census of the objects we find. Astronomical surveys have always been a key aspect of our field: such surveys have much to teach us about the formation and structure of the Milky Way galaxy in which the sun sits; the expansion, future fate, and origin of the universe as a whole; and the nature of stars and the planets that orbit them. Indeed, astronomy is the study of origins; we ask (and occasionally answer!) the most fundamental questions about where planets, stars, galaxies, and the universe as a whole come from. With each advance in technology and new way to survey the heavens, we uncover new phenomena that we did not anticipate, and we find ourselves addressing questions that we previously did not have the imagination to ask.

With one hundred billion stars in our Milky Way, and one hundred billion galaxies in the observable universe, our surveys have entered the realm of big data. The biggest survey telescopes today can gather terabytes of data in a single night, and our catalogs of galaxies and stars include over a billion objects, a number that will increase by a factor of ten over the next decade or so. The discovery potential of our surveys is limited by a combination of raw computer processing power, the cleverness of our algorithms, and our imagination. Just as Google allows us to query human databases to uncover facts and the relations between them, astronomers have been developing similar technology to query the database of the universe.

This essay will focus on surveys in visible light, which of course represents only a small sliver of the full range of electromagnetic waves, from high-energy gamma rays to long-wavelength radio waves. Very different physical phenomena are responsible for emission at different regions of the electromagnetic spectrum, and comparison of maps of the sky in these different regimes is a powerful tool for exploring these phenomena. However, normal stars (and thus galaxies, which are made up of stars) emit most of their radiation at visible and near-infrared wavelengths, making this the most effective regime in which to survey the sky for these objects.

Go outside on a clear moonless night, far from the lights of civilization. Once your eyes have adapted to the darkness, between two and three thousand stars are discernable to the naked eye at any given time. You will also see a silvery band crossing the sky: the Milky Way. Galileo Galilei was the first to point a telescope to the heavens, and he discovered that the light of the Milky Way comes from countless stars. Since that time, astronomers have used ever larger telescopes to map objects in the sky.

However, the images we make with cameras placed at the back of our telescopes are two-dimensional. We have no depth perception, and the stars look all to be equidistant, with no sense of which are closer and which are farther away. In fact, the nearest star (other than the sun) is about four light years (or about forty trillion kilometers) from us, a distance that is completely outside our everyday experience. It would take thirty thousand years to cover that distance traveling at the speed of our fastest spacecraft (forty-five kilometers per second). While most of the individual stars visible to the naked eye are within a few hundred light years of us, the bulk of stars in the Milky Way are much farther away, arrayed in a vast flattened spiral structure some one hundred thousand light years across, containing roughly one hundred billion stars.

One hundred years ago, astronomers understood the Milky Way galaxy to be the full extent of the universe. However, in addition to the myriad stars apparent in astronomical images, one also sees fuzzy extended objects, termed *nebulae* (the Latin word for *cloud*). Edwin Hubble demonstrated in the 1920s that these nebulae were other "island universes," as large as our own Milky Way but at much greater distances. This discovery enormously expanded our understanding of the size of the universe. The nearest big galaxy to our own is about two million light years distant, and the number of galaxies seen in the deepest images with the eponymous Hubble Space Telescope imply that there are about one hundred billion of them in the observable universe.

Surveys of galaxies based on photographic plates show that they are far from uniformly distributed in the sky: clusters a few million light years across containing hundreds of galaxies are apparent, with hints of larger structures yet. But to really map the distribution of galaxies in three dimensions, we need an unambiguous way to measure their distances. Hubble's second great discovery – that the universe is expanding – gives us the way to do so.

Consider the *spectrum* of an astronomical object, which measures the intensity of its light as a function of wavelength. This spectrum gives much more detailed information about the physical nature of the object than the properties (size, brightness, and color) measureable from an astronomical image. For example, the wavelength at which the spectrum of a star peaks is a measure of its surface temperature, which for most stars in turn indicates their mass. As described in Anna Frebel's article in this volume, there are characteristic wavelengths at which the star is dimmer (*absorption lines*); these are due to absorption of light by atoms in the atmosphere of the star, and they reflect the star's chemical composition.

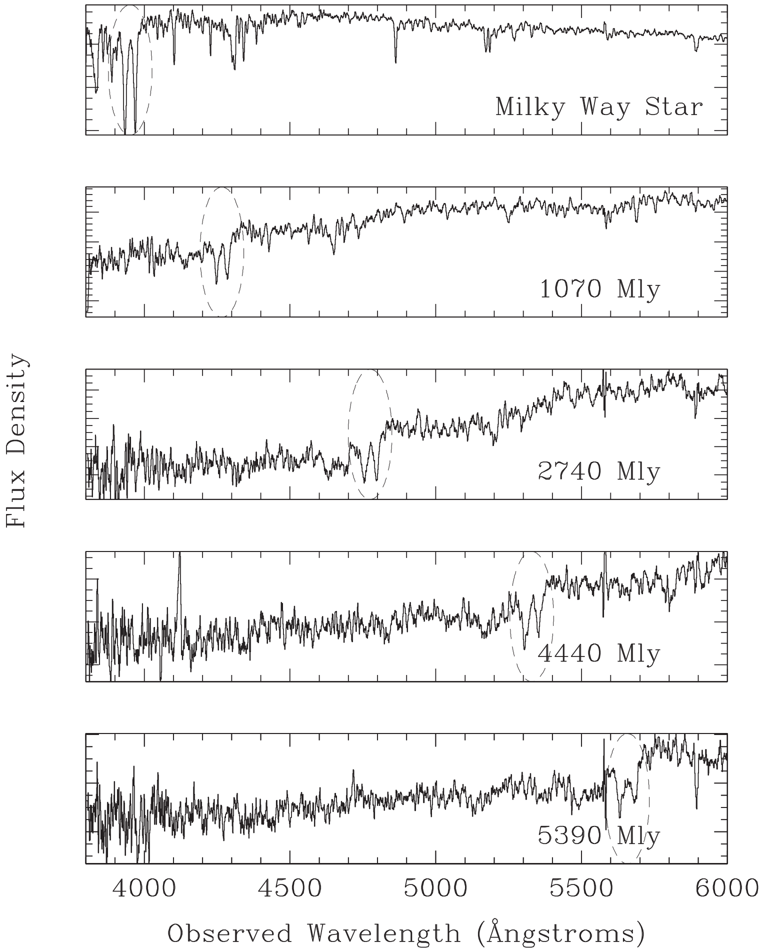
As illustrated in Figure 1, the spectra of galaxies show many of the same absorption lines as do stars like the sun, which tells us that galaxies are made of stars. However, there is an important difference: the absorption lines in galaxy spectra are shifted systematically to longer (redder) wavelengths. Hubble found that the spectra of essentially all galaxies are redshifted, and the degree of redshift is proportional to the distance of the galaxy.

This relationship between redshift and distance is a consequence of the expansion of the universe and, as explained in David Spergel's companion article in this volume, it leads directly to our modern understanding of the Big Bang. For our purposes, however, this becomes a valuable tool for mapping the universe in three dimensions: measuring the spectrum of a galaxy allows us to determine its redshift, and thus, by Hubble's law, its distance.

Photographic film records the presence of only (at best) a few percent of the photons that fall upon it. Thus, even with the largest telescopes available in Hubble's time, measuring the spectrum of a faint galaxy in order to determine its redshift was enormously time-consuming, requiring exposure times of many hours for even the nearest galaxies. Modern electronic detectors, such as those in your digital camera, are far more sensitive, detecting close to 100 percent of the photons that fall on them. Their development and adoption by the astronomical community starting in the late 1970s meant that appreciable numbers of galaxy spectra, and thus redshifts and distances, could be measured.

The galaxy distribution based on these early surveys of a few thousand redshifts were stunning and surprising. These maps showed that most galaxies are strung along long filamentary structures that connect the rich clusters. These filaments, up to hundreds of millions of light years across, surround enormous volumes, essentially devoid of galaxies. These pictures raised many questions: How large can these

Figure 1
Spectra of Objects from the Sloan Digital Sky Survey



The upper panel shows a star similar to the sun; the pair of dips (circled) at a wavelength just below 4000 A is due to absorption by calcium atoms in the atmosphere. The other panels show spectra of galaxies at ever greater distances (marked in each panel in units of millions of light years, or Mly). The calcium absorption feature is circled in each case; the more distant the galaxy, the larger the redshift. Source: The Sloan Digital Sky Survey, http://www.sdss.org/.

structures be? How did they form? What can they tell us both about the nature of galaxies and the structure of the universe overall?

All this motivated the next generation of redshift surveys, the most ambitious of which was the Sloan Digital Sky Survey (SDSS).[1] This program was the scientific vision of Princeton University astronomer James Gunn, who realized in the late 1980s that advances in electronic detector technology, telescope design, and computer processing power made it possible to design a dedicated telescope project to map the universe much more thoroughly than had been done to date. The initial stated goal of the project was to measure redshifts for one million galaxies in the local universe in order to characterize the nature of the galaxy distribution on the largest scales. The SDSS built a specialized telescope with a two-and-a-half-meter-diameter primary mirror, as well as an astronomical camera with a total of 145 megapixels (the largest ever at the time it was built). The telescope swept the sky essentially every clear moonless night for ten years, covering twenty square degrees every hour in five broad filters. The resulting color images, which survey the region of sky far from the band of the Milky Way (whose dust obscures the light of more distant galaxies) now cover about one-third of the celestial sphere and contain data on almost half a billion stars, galaxies, quasars, and asteroids.

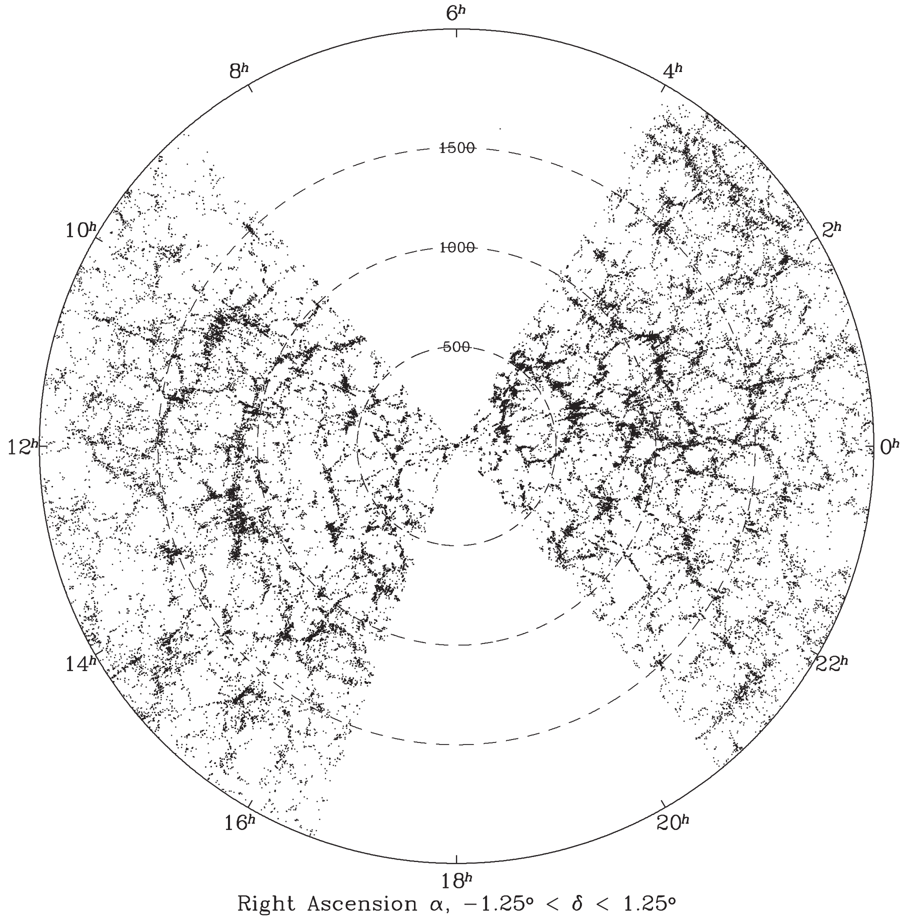
On the less than pristine nights, the imaging camera was replaced with a spectrograph, fed by optical fibers from the focal plane of the telescope, allowing spectra of 640 objects selected from the images (increased to 1000 objects after a major upgrade to the system in 2009) to be measured at a time. To date, the survey has measured the spectra, and thus redshifts and distances, of over two million galaxies, creating a stunning and detailed map of the universe in which we live. Figure 2 covers only a few percent of the full sample; each of the roughly fifty thousand dots shown represents an entire galaxy, as large as the Milky Way, containing one hundred billion stars. The filaments and voids apparent in the first redshift surveys of the early 1980s appear in all their glory here. The largest structure in the map, dubbed "The Sloan Great Wall," is 1.4 billion light years across. Interestingly, there appears to be an "end to greatness," to use astrophysicist Bob Kirshner's memorable phrase: there are no structures even larger than this; and if we step back far enough, the distribution of galaxies appears uniform.

As Spergel's article describes, these data, together with measurements of the cosmic microwave background and other cosmological probes, lead to a comprehensive model for the structure and makeup of the universe. In particular, we can understand the richness of the distribution of galaxies in a model in which ordinary matter (in the form of atoms, mostly hydrogen, which make up the stars that cause galaxies to shine) represents just under 5 percent of the mass-energy density of the universe, with the remainder being in the form of dark matter and dark energy.

One lesson of astronomy surveys is that by gathering data necessary to answer one question (in this case, wide-field imaging and spectroscopy to measure the large-scale distribution of galaxies), astronomers gain the ability to address other problems in the field, many of which were unanticipated at the beginning of the survey. The SDSS is no exception to this rule, making fundamental discoveries in the structure of the Milky Way galaxy, the most distant quasars and the nearest stars, and many other topics.[2]

About half the objects visible in the SDSS images are stars in our own Milky Way, at typical distances of thousands to tens of thousands of light years, appearing as sharp points of light; most of the rest are galaxies, which appear fuzzy in the images. Away from the densest part of the Milky Way, a flattened rotating disk, stars lie in what is thought to be a roughly spherical distribution around the center of our galaxy, termed the *halo* of the Milky Way, extending to tens of thousands of light years. It was long thought that the halo was smooth, with stellar density falling steadily as one moves from the galactic center. But the maps of the stars from the

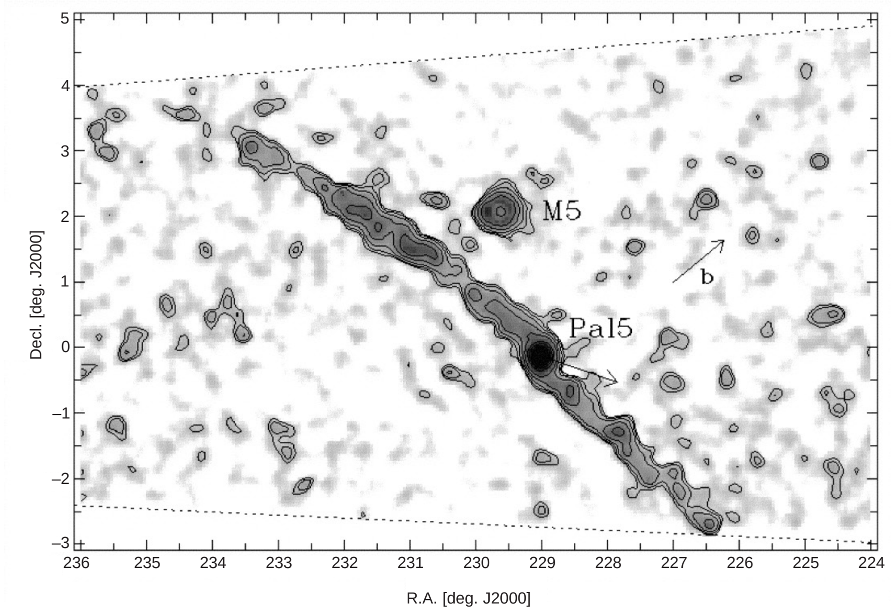
Figure 2
A Slice through the Three-Dimensional Distribution of Galaxies Mapped by the Sloan Digital Sky Survey



Each of the more than fifty thousand dots in this figure represents a galaxy as large as the Milky Way. The Milky Way itself sits in the center of the figure; the outer circle represents a distance of two billion light years from the sun. The filamentary nature of the galaxy distribution is readily apparent in this figure. The segments devoid of data are regions of the sky that the SDSS did not survey. Source: The Sloan Digital Sky Survey, http://www.sdss.org/.

SDSS and from earlier photographic surveys show that things are more interesting than that. A taste of the results is shown in Figure 3, which maps the distribution of stars in the vicinity of a so-called globular cluster (named Palomar 5), a conglomeration of roughly one hundred thousand stars 150 light years across. Palomar 5 is accompanied by a stream of stars more than ten thousand light years long. It is now understood that as the globular cluster orbits our Milky Way, gravitational tidal forces (the difference in the gravitational acceleration between the near and far sides of the globular cluster) are tearing the cluster apart, pulling streams of stars from it. Indeed, the star maps from the SDSS have shown that such star streams are com-

Figure 3
The Distribution of Stars in the Vicinity of the Globular Cluster Palomar 5



The globular cluster itself is the dark region to the lower-right of the center of the figure. It is accompanied by both leading and trailing streams of stars, stretching over ten thousand light years. This stream is believed to have been pulled out from the globular cluster by tidal forces from the Milky Way. Source: Michael Odenkirchen et al., "The Extended Tails of Palomar 5: A 10° Arc of Globular Cluster Tidal Debris," *The Astronomical Journal* 126 (2003): 2385. Reprinted with permission.

mon: in a process predicted by some theorists and now confirmed by these observations, the halo appears to be made largely of the debris of small galaxies and globular clusters that have fallen into, and been torn apart by, the gravity of our Milky Way.[3]

Surveying large swaths of sky to very faint levels makes one sensitive to very rare (and therefore interesting) objects, such as quasars at very great distances. Because of the finite speed of light, we see a distant galaxy not as it is today, but as it was when the universe was significantly younger. Astronomers thus can use their telescopes as time machines to directly observe the history of the universe. Of course, very distant galaxies will be tremendously faint as seen from Earth, so, all else being equal, the easiest distant galaxies to see will be those with the highest intrinsic luminosities. These include quasars: galaxies with a supermassive black hole (a black hole with a mass up to several billion times that of the sun) in their center; as described in Scott Tremaine's article in this volume, gas orbiting close to the black hole heats up tremendously and glows enough to outshine the hundred billion stars of the galaxy by a factor of one hundred or more.

Because the light from a quasar is dominated by the tiny region in the vicinity of the black hole, they are usually unresolved in SDSS images, and thus look like stars. The SDSS was designed to identify quasars by their distinctive colors in the images. In particular, due to absorption by neutral hydrogen in their spectra, the highest-redshift quasars are extremely red, appearing only in the longest-wavelength filter that the SDSS measures (the "z" band). Soon after the SDSS had garnered its first images of the sky, my student Xiaohui Fan (now a professor at the University of Arizona) and I started searching for record-breaking quasars.

Finding these most distant quasars is straightforward in principle: simply identify those objects that appear only in the z band, and confirm that they are quasars by taking a spectrum. Making this work in practice, however, required a very detailed understanding of every way the data could fool us; even extremely rare glitches that affect one in a million stars would swamp our search for these objects. We learned that doing science from surveys requires exquisite quality control and a deep understanding of the nature of the data!

But we were stymied in our search by another problem, this one astrophysical. Extremely cool stars also appear quite red. About the time that the SDSS was taking its first data, astronomers were using near-infrared surveys of the sky to discover new classes of very cool red stars, called brown dwarfs. These have very low masses, so low that their gravity is inadequate to ignite thermonuclear fusion in their cores. Indeed, these are only a bit more massive than many of the planets described in Gáspár Bakos's article in this volume on exoplanets. Brown dwarfs have surface temperatures below 2000 K (for comparison, the sun has a surface temperature of 6000 K), making them dim (and thus only visible at very small distances from Earth, typically thirty light years or less) and very red, just like the quasars. In our search for the most distant quasars in the universe, we found ourselves stumbling over some of the nearest stars to our own!

Despite these problems (and the serendipity of finding brown dwarfs was exciting and scientifically important in its own right), we broke the record multiple times for the most distant quasar discovered, including one whose light we observe today was emitted less than a billion years after the Big Bang. This is remarkable: the cosmic microwave background, emitted by gas in the universe four hundred thousand years after the Big Bang is smooth to one part in one hundred thousand, and yet a billion years later, a supermassive black hole, the densest conceivable object with a mass billions of times that of the sun, had managed to form. It is still poorly understood how this happened, and discoveries of yet more distant quasars may shed some light into this process. Our record has since been broken in spectacular fashion with a quasar seen one hundred million years earlier than our own, found in data from the United Kingdom Infrared Telescope Deep Sky Survey.

One of the most important lessons from the SDSS and other surveys is that if the data are of high quality and are made available to the world, they enable science far beyond the initial goals of the survey. The SDSS, which is now in its fifteenth year of full operation, has plans to continue gathering data through the year 2020. Almost six thousand refereed articles have been written to date by scientists all over the world using the SDSS data, a level of productivity that rivals that of much larger telescopes, such as the ten-meter Keck Telescopes in Hawaii. Thus astronomers are motivated to consider the next generation of surveys beyond the SDSS. The SDSS telescope has a primary mirror with a diameter of only two-and-a-half meters, which is small compared to the largest optical telescopes in the world today: ten meters across (and three significantly larger telescopes are being planned for the next decade, with diameters of twenty-two, thirty, and thirty-nine meters). Larger telescopes are capable of seeing much fainter, and therefore more distant, galaxies. As we have already seen, this means that they can probe back to times when the galaxies, and the universe as a whole, were much younger than they are today.

As Pieter van Dokkum's article in this volume describes, the early universe was very different from the universe today. The first galaxies are thought to have formed several hundred million years after the Big Bang, and one of the most important areas of research today is to determine how they grew and evolved from their initial formation. The next generation of planned surveys is designed to address this question. The 8.2-meter Subaru Telescope operated by the National Astronomical Observatory of Japan has the largest field of view of any existing telescope of its size, and is thus particularly well-suited for carrying out surveys. It has a wide-field imaging camera that has just started a major survey covering over one thousand square degrees of sky, sensitive to objects twenty-five times fainter than the SDSS was able to reach. In 2018, this will be followed up with a spectroscopic survey of hundreds of thousands of galaxies from the first few billion years after the Big Bang.

The era of surveys will continue into the next decade. The Large Synoptic Survey Telescope (LSST) will spend ten years in a dedicated wide-field imaging survey of the sky.[4] Polishing of its 8.4-meter primary mirror is nearing completion; the telescope will be constructed in the Chilean Andes (famed for their clear and steady skies) and will start a ten-year survey of the sky in 2022, covering half the celestial sphere four times deeper than the Subaru survey will go. In doing so, it will map the heavens multiple times, making a movie of the sky: "celestial cinematography," as LSST Chief Scientist Tony Tyson likes to phrase it. Indeed, surveys on much smaller telescopes that repeatedly image the sky have discovered a wide range of variable phenomena, including the motions of asteroids in our own solar system, pulsating and exploding stars of all sorts, the subtle motions of distant stars that reflect the gravitational potential of the Milky Way, and the flickering of quasars as parcels of gas swirl around the black holes that power them. The LSST will extend such studies to much fainter astronomical objects, and is bound to find new kinds of variable and transient phenomena that have not been anticipated to date.

Telescope observations from the surface of the Earth are affected by the atmosphere, which both blurs images and adds a substantial amount of background light, especially at near-infrared wavelength. As the Hubble Space Telescope has dramatically demonstrated, placing an observatory in space allows much sharper images and the ability to observe much fainter objects. The Hubble Telescope itself has a tiny field of view, and is thus not well-suited for survey work, but the U.S. National Reconnaissance Office has recently given NASA two telescopes that have mirrors as large as Hubble's (2.4 meters in diameter, similar to that of the SDSS) but are designed with a much larger field of view. There are active plans to use one of them to map the large-scale distribution of dark matter by measuring the distortions it causes on the shapes of galaxies. The combination of this space-based telescope with the LSST will be particularly powerful; the coarser images of the LSST will be balanced by its much larger sky coverage and measurements over a wide range of wavelengths. The quantity of data these surveys will produce will be measured in petabytes (one petabyte is a million gigabytes); analyzing and interpreting these data will drive new technologies in computer processing and analysis. These data will be made public, allowing schoolchildren to search for supernovae at the same time as professional astronomers. With these surveys well underway a decade from now, astronomical discovery will continue to be driven by the amazing datasets that they produce.

ENDNOTES

[1] The Sloan Digital Sky Survey is described in detail at http://www.sdss.org. See also the popular book describing the building of the survey: Ann Finkbeiner, *A Grand and Bold Thing: An Extraordinary New Map of the Universe Ushering in a New Era of Discovery* (New York: Free Press, 2010).

[2] The discovery of distant quasars in the SDSS is described in Xiaohui Fan et al., "A Survey of z > 5.7 Quasars in the Sloan Digital Sky Survey," *The Astronomical Journal* 125 (2003): 1649.

[3] The structure of the halo of the Milky Way as seen in the SDSS is described in V. Belokurov et al., "The Field of Streams: Sagittarius and Its Siblings," *The Astrophysical Journal* 642 (2006): L147.

[4] The Large Synoptic Survey Telescope is described at http://www.lsst.org. A comprehensive description of its science opportunities may be found in the *LSST Science Book v.* 2.0 (2009) at http://www.lsst.org/lsst/scibook.

The Odd Couple: Quasars & Black Holes

Scott Tremaine

Abstract: Quasars emit more energy than any other object in the universe, yet are not much bigger than our solar system. Quasars are powered by giant black holes of up to ten billion (10^{10}) times the mass of the sun. Their enormous luminosities are the result of frictional forces acting upon matter as it spirals toward the black hole, heating the gas until it glows. We also believe that black holes of one million to ten billion solar masses – dead quasars – are present at the centers of most galaxies, including our own. The mass of the central black hole appears to be closely related to other properties of its host galaxy, such as the total mass in stars, but the origin of this relation and the role that black holes play in the formation of galaxies are still mysteries.

SCOTT TREMAINE, a Foreign Honorary Member of the American Academy since 1992, is the Richard Black Professor in the School of Natural Sciences at the Institute for Advanced Study and the Charles A. Young Professor of Astronomy on the Class of 1897 Foundation, Emeritus at Princeton University. His research interests are centered on astrophysical dynamics, including planets, small bodies in the solar system, galaxies, black holes, and galactic nuclei. His work is published in such journals as *The Astrophysical Journal* and *Monthly Notices of the Royal Astronomical Society*. With James Binney, he is the author of the graduate textbook *Galactic Dynamics*.

Black holes are among the most alien predictions of Einstein's general theory of relativity: regions of space-time in which gravity is so strong that nothing – not even light – can escape. More precisely, a black hole is a singularity in space-time surrounded by an event horizon, a surface that acts as a perfect one-way membrane: matter and radiation can enter the event horizon, but, once inside, can never escape. Although black holes are an inevitable consequence of Einstein's theory, their main properties were only understood – indeed, the name was only coined – a half-century after Einstein's work.[1] Remarkably, an isolated, uncharged black hole is completely characterized by only two parameters: its mass and its spin (or angular momentum). An eloquent tribute to the austere mathematical beauty of these objects is given by the astrophysicist Subrahmanyan Chandrasekhar in the prologue to his monograph *The Mathematical Theory of Black Holes*: "The black holes of nature are the most perfect macroscopic objects there are in the universe: the only elements in their construction are our concepts of space and time. And since the general theory of relativity provides only a single unique

doi:10.1162/DAED_a_00310

family of solutions for their descriptions, they are the simplest objects as well."[2] (Anyone who scans the six hundred pages that follow, however, is unlikely to agree that they are as simple as claimed.) Simple or complex, we are now almost certain, for reasons outlined in this essay, that black holes do exist and that a giant black hole of several million times the mass of the sun is present at the center of our galaxy.

Laboratory study of a black hole is impossible with current or foreseeable technology, so the only way to test these predictions of Einstein's theory is to find black holes in the heavens. Not surprisingly, isolated black holes are difficult to see. Not only are they black, they are also very small: a black hole with the mass of the sun is only a few kilometers in diameter (this statement is deliberately vague: because black holes bend space to such extremes, our notions of "distance" close to a black hole cease to be unique). However, the prospects for detecting black holes in gas-rich environments are much better. The gas close to the black hole normally takes the form of a rotating disk, called an accretion disk: rather than falling directly into the black hole, the orbiting gas gradually spirals in toward the event horizon as it loses orbital energy, most likely transferred to turbulence and magnetic fields in the disk.[3] The energy is eventually transformed into thermal energy, which heats the gas until it begins to glow, mostly at ultraviolet and X-ray wavelengths. By the time the inward-spiraling gas disappears behind the event horizon, deep within the gravitational well of the black hole, a vast amount of radiation has been emitted from every kilogram of accreted gas.

In this process, the black hole can be thought of as a furnace: when provided with fuel (the inward-spiraling gas) it produces energy (the outgoing radiation). Einstein's iconic formula $E = Mc^2$ relates mass M and the speed of light c to an energy E called the rest-mass energy. Using this relation, there is a natural measure of the efficiency of this or any other furnace: the ratio of the energy it produces to the rest-mass energy of the fuel that it consumes. For furnaces that burn fossil fuels, the efficiency is extraordinarily small (about 5×10^{-10}), and all combustion processes based on chemical reactions have similarly low efficiencies. For fission reactors using uranium fuel, the efficiency is much better, around 0.1 percent; and for the fusion reactions that power the sun and stars, the efficiency can reach 0.3 percent.

Black-hole furnaces can have far higher efficiency than any of these: between 10 and 40 percent for accretion of gas from a thin accretion disk. In the unlikely event that we could ever domesticate black holes, the entire electrical energy consumption in the United States could be provided by a black-hole furnace consuming only a few kilograms of fuel per year.

Despite the relatively low efficiency of fusion reactions, most of the light in the universe comes from stars. Most of the stars in the universe are organized in galaxies: assemblies of up to 10^{11} stars orbiting in a complex dance determined by their mutual gravitational attraction. Our own galaxy contains a few tens of billions of stars arranged in a disk; the nearest of these is about 1 parsec (3.26 light years) from us, and the distance to the center of our galaxy is about 8 kiloparsecs (or about 26,000 light years). The diffuse light from distant stars in the galactic disk is what we observe as the Milky Way.[4]

A small fraction of galaxies contain mysterious bright and compact sources of radiation at or near their centers, called active galactic nuclei.[5] The brightest of these are the quasars; remarkably, they can emit up to 10^{13} times more light than stars like the sun, thereby outshining the entire galaxy that hosts them. Even though quasars are much rarer than galaxies, they are so bright that they contribute almost 10 percent of the light emitted in the universe.

Ironically, the extraordinary luminosity of quasars is what made them hard to discover. Except in a few cases, they are so bright that the host galaxy cannot be seen in the glare from the quasar, and so small that they look like stars, even at the resolution of the Hubble Telescope (in fact, "quasar" is a contraction of "quasi-stellar object"). Thus, even the brightest quasars are usually indistinguishable from millions of stars of similar brightness. Fortunately, some quasars are also strong sources of radio emission, and in 1963 this clue enabled astronomer Maarten Schmidt at Caltech to identify a radio source called 3C 273 with a faint optical source that otherwise looked like an undistinguished star.[6] With this identification in hand, Schmidt was able to show that 3C 273's spectral lines were redshifted – Doppler shifted by the cosmological expansion of the universe – to wavelengths 16 percent longer than laboratory spectra, and thus that 3C 273 was at a distance of eight hundred megaparsecs, ten million times farther away than it would have been if it were a normal star.

Now almost one hundred thousand quasars have been identified. The most distant of them is almost one hundred times farther away from Earth than 3C 273, and its light was emitted when the universe was only 6 percent of its present age. However, the formation of quasars that early in the history of the universe was a very rare event. Most were formed when the universe was 20–30 percent of its current age, and quasars today are a threatened species: the population has declined from its peak by almost two orders of magnitude, presumably because the fuel supply for quasars dried up as the universe expanded at an accelerating rate.

How can quasars emit so much energy? The suggestion that they are black-hole furnaces was made independently by Edwin Salpeter in the United States and Yakov Zel'dovich in the Soviet Union, soon after quasars were first discovered. But in the 1960s the black hole was a novel and exotic concept, and staggeringly massive black holes (roughly one hundred million solar masses) were required in order to explain quasar properties. Thus, most astronomers quite properly focused on more conservative models for quasars, such as supermassive stars, dense clusters of ordinary stars or neutron stars, and collapsing gas clouds. Over the next two decades, however, all of these models were subjected to intense scrutiny that increasingly suggested that they were inadequate to explain the growing body of observations of quasars. Furthermore, other studies showed that the alternative models generically evolve into a black hole containing most of the mass of the original system, thereby suggesting that the formation of massive black holes is natural and perhaps even inevitable.

A number of indirect but compelling arguments also support the black-hole furnace hypothesis. The first of these relates to efficiency. The luminous output of a bright quasar over its lifetime corresponds to a rest-mass energy of about one hundred million times the mass of the sun. If this were produced by the fusion reactions that power stars with the efficiency of 0.3 percent given earlier, the mass of fuel required would be almost the total of all the stars in our galaxy. There is no plausible way to funnel this much mass into the tiny central region that the quasar occupies, nor is there evidence that so much mass resides there. On the other hand, for a black-hole furnace the efficiency is 10 percent or more, so the required mass is less than 10^9 solar masses, and this much gas is not hard to find close to the center of many galaxies. Thus, the black-hole furnace is the only model that does not bankrupt the host galaxy's fuel budget.

A second argument is based on the small size of quasars. We have known since their discovery that quasars appear as unresolved point sources of light even in the best optical telescopes; this observation alone implies that they must be less than about a kiloparsec across, and long-baseline radio interferometry shows that quasars must be far smaller, less than one parsec across. Even stronger limits on the size come from indirect measurements. Quasars vary irregularly in brightness on a wide range of timescales, from weeks to decades and probably even longer. It proves quite difficult to construct any plausible model of a luminous astrophysical object that varies strongly on a timescale smaller than the time it takes light to travel across the object: the separate parts of the object are not causally connected on this timescale, so they vary independently and their contributions tend to average out. This argument suggests that the size of the most rapidly varying quasars must be less than the distance light travels in a few weeks (which is around a few hundredths of a parsec or a few thousand times the Earth-sun distance). This upper limit is consistent with size estimates from a number of other methods, such as reverberation mapping, photoionization models, and gravitational lensing.[7]

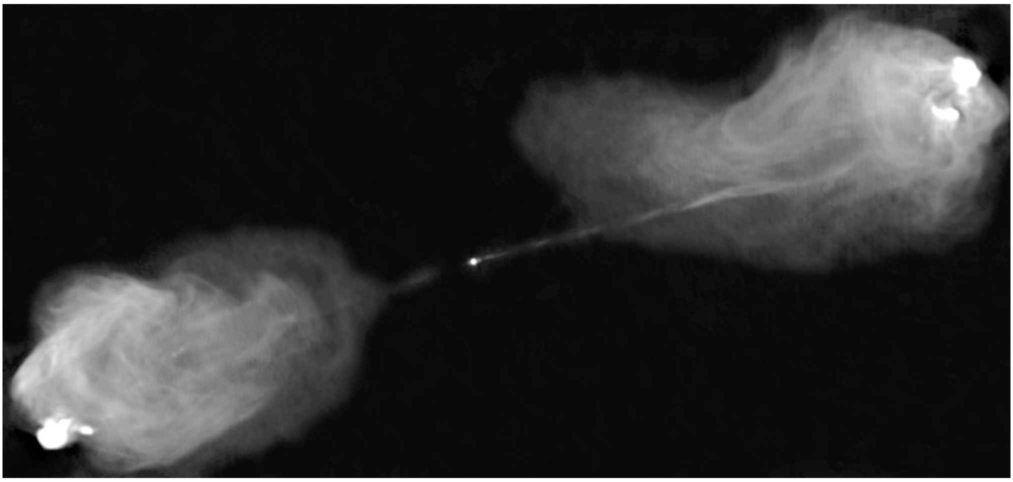
A few hundredths of a parsec is large by our standards but extremely small on galactic scales: a millionth of the size of the galaxy as a whole. A black hole of one hundred million solar masses and its surrounding accretion disk would fit comfortably inside this volume – its relativistic event horizon has a radius of about the Earth-sun distance – but almost all of the alternative models that might explain quasars fail to do so.

In a few cases space-based observatories can measure X-ray spectral lines emitted by quasars. These are not the narrow lines seen in spectra of the sun, stars, or interstellar gas; instead they are grossly misshapen, with broad tails extending to much longer wavelengths than such lines would have in weak gravitational fields. The only plausible explanation for these distortions is that they arise from gravitational redshift – the loss of energy as the X-rays climb out of a deep gravitational well on their way to us – and/or from extreme Doppler shifts caused by relativistic motions, most probably in a rotating accretion disk. Either explanation requires that the X-rays were emitted from a region only a few times larger than the event horizon of a black hole, as no other known astrophysical system has such high velocities and deep potential wells.[8]

Some quasars emit powerful jets of plasma that extend for up to a megaparsec (see Figure 1),[9] probably collimated and accelerated by magnetic fields near the black hole that are twisted up by the rotation of the surrounding accretion disk. The production of these jets is not so remarkable: for example, various kinds of star also produce jets, though on a much smaller scale. What is more striking is that quasar jets typically travel at close to the speed of light. Once again, there is no plausible way to produce such high velocities except close to the event horizon of a black hole. Moreover, in most cases the jets are accurately straight, even though the innermost plasma in the jet was emitted a million years after the material at the far end. Thus, whatever mechanism collimated the jet must maintain its alignment over several million years; this is easy to do if the jets are squirted out along the polar axis of a spinning black hole, but difficult or impossible in other quasar models.

Finally, there is strong evidence that a handful of systems that emit strong X-ray

Figure 1
An Image at Radio Wavelengths of Jets from the Quasar Cygnus A



The bright spot at the center of the image is the quasar, which is located in a galaxy 240 megaparsecs away. The long, thin jets emanating from the quasar terminate in bright "hotspots" when they impact the intergalactic gas that surrounds the galaxy. The hotspots are roughly 70 kiloparsecs (or 228,000 light years) from the quasar. The brighter of the two jets is traveling toward us; its brightness has been boosted by relativistic effects. Source: National Radio Astronomy Observatory/Associated Universities Inc.; reproduced by permission of the American Astronomical Society. From R. A. Perley et al., "The Jet and Filaments in Cygnus A," *The Astrophysical Journal* 285 (1984): L35 – L38.

radiation consist of a normal star and a black hole. The black holes are much less massive than those in quasars, only a few times the mass of the sun.[10] The black hole and the normal star orbit one another at a small enough distance – a few stellar radii – that material lost from the normal star fuels a miniature black-hole furnace. These systems reinforce our confidence in the existence of black holes, and allow us to refine our understanding of the complex physics of a black-hole furnace.

Based on these and other arguments, there is near-complete agreement among astrophysicists that the power source for quasars is the accretion of gas onto black holes of one hundred million solar masses or more. Accepting this position leads to a simple syllogism that has driven much of the research on this subject for the past several decades: if quasars are found in galaxies, and the number of quasars shining now is far smaller than when the universe was young, and quasars are black-hole furnaces, then many "normal" galaxies should still contain the massive black holes that used to power quasars at their centers, but are now dark. Can we therefore find "dead quasars" in nearby galaxies?

There are two important guideposts in the search for dead quasars. The first comes from a simple argument by the Polish astronomer Andrzej Sołtan.[11] We know that the universe is homogeneous on large scales, and therefore on average the energy density in quasar light must be the same everywhere in the universe (here *average* means averaged over scales greater than about ten to twenty megaparsecs, which is still small compared to the overall "size" of the universe, at a few thousand megaparsecs). We can measure this energy density by adding up the contributions from all the quasars found in surveys (after straightforward corrections for incompleteness). If this energy were produced by black-hole furnaces with an efficiency of 10 percent, for example, then a mass M of material accreted by black holes would produce $0.1Mc^2$ in quasar light. Similarly, if the average mass density of dead quasars is ρ, then the energy density of quasar light must be $0.1\rho c^2$.

Since we know the latter figure, we can invert the calculation to determine the mass density of dead quasars. The power of this argument is that it requires no assumptions about the masses or numbers of black holes; no knowledge of when, where, or how quasars formed; and no understanding of the physics of the quasar furnace except its efficiency. Sołtan's argument tells us that the density of dead quasars should be a few hundred thousand solar masses per cubic megaparsec, compared to a density of large galaxies of about one per hundred cubic megaparsecs. What it does not tell us is how common dead quasars are: on average there could be, for example, one dead quasar of ten million solar masses in every galaxy, or one of one billion solar masses in 1 percent of galaxies.

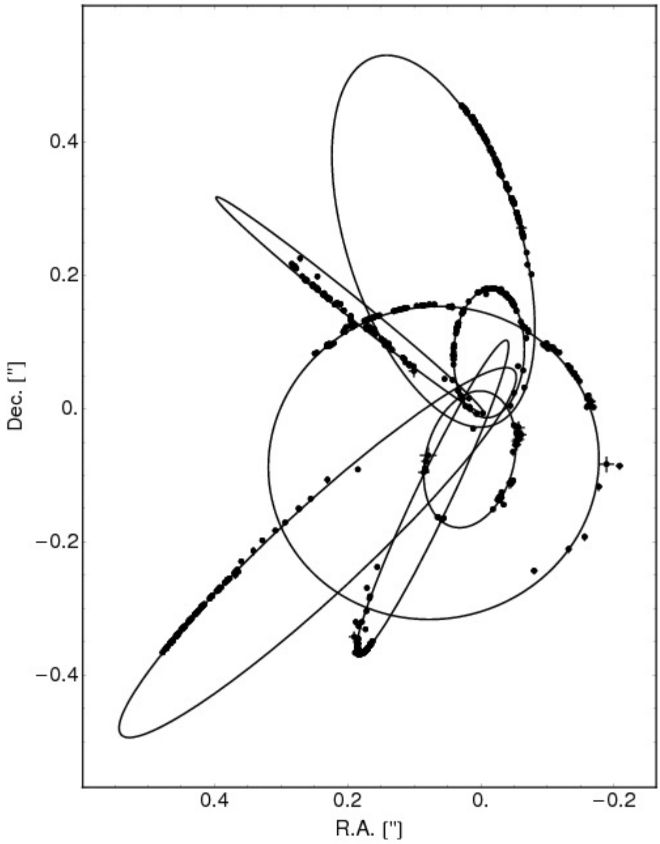
The second guidepost is that the centers of galaxies are the best places to prospect for dead quasars. There are several reasons for this. First, live quasars seem to be found near the centers of their host galaxies (although this is difficult to tell with precision because the glare from the quasar obscures the structure of the host). Second, the fuel supply for a black hole sitting at rest in the center of the galaxy is likely to be much larger than the fuel supply for one orbiting in the outskirts of the galaxy. Third, massive black holes orbiting in a galaxy tend to lose orbital energy through gravitational interactions with passing stars, so they spiral into the center of the galaxy.[12] And finally, like the drunkard looking for his keys under the lamppost, we look for dead quasars at the centers of galaxies because that is where it is easiest to find them: the search area is small and the density of stars that are affected by the black hole's gravitational field is high.

Stars that come under the influence of the black hole's gravitational field – typically those within a distance of a few tenths of a parsec to a few tens of parsecs, depending on the black hole's mass – are accelerated to higher velocities; although individual stars cannot be detected in galaxies other than our own, this acceleration leads to increased Doppler shifts, which broaden the spectral lines from the collective stellar population. This broadening can be detected by spectroscopic observations with sufficiently high spatial resolution and signal-to-noise ratios. The search for this effect in the centers of nearby galaxies began around 1980 and yielded evidence for black holes in a handful of cases. Strictly, the evidence was for massive dark objects with masses of millions to billions of solar masses, since the angular resolution (the smallest size of distinct object that the telescope can clearly image) of these observations was still far larger than the size of the event horizon of the putative black hole. These results were tantalizing, but incomplete: the problem was that the angular resolution of ground-based telescopes is limited by blurring caused by the atmosphere, so the effects of a black hole could be detected only in the closest galaxies, and then only over a limited range of distances from the center. Precisely this problem was one of the motivations for constructing the Hubble Space Telescope, which at the time of its launch in 1990 had roughly ten times the angular resolution of the best ground-based telescopes. Since then the Hubble Telescope has devoted many hundreds of hours to the hunt for black holes at the centers of galaxies, and by now Hubble has confirmed and strengthened the ground-based detections in nearby galaxies and produced firm evidence for black holes in over two dozen more distant ones.[13] Even in the best cases, this method can only probe to a few tenths of a parsec from the galaxy center, but we are persuaded that the massive dark objects observed by Hubble must be black holes because the alternatives (for example, a cluster of low-luminosity stars) are far less plausible. By now the Hubble Telescope has turned to other tasks, but the search for dead quasars has been resumed by ground-based telescopes, now using adaptive optics systems that can correct for atmospheric blurring in real time. Adaptive optics is beginning to provide angular resolutions that equal or exceed Hubble's: these new observations also have far higher signal-to-noise ratios, since the collecting area of the biggest ground telescopes is ten times that of Hubble.

The painstaking measurement of stellar motions near the centers of galaxies has been supplemented by an unexpected gift from the heavens: the otherwise unremarkable galaxy NGC 4258 contains at its center a thin, nearly flat, rotating disk of gas, about a tenth of a parsec in radius. The gas includes water vapor, and the temperature and density in the disk are right for the production of maser (microwave laser) emission in the water, stimulated by a weak active galactic nucleus at the center of the disk. The maser emission consists of tiny, intensely bright sources of radiation concentrated in wavelength at the spectral line of water, and by measuring the Doppler shift of these sources and their motion across the sky using an intercontinental array of radio telescopes, we can map out the rotation of the disk with exquisite precision. The disk is found to rotate around the active nucleus; from the disk kinematics, we can deduce that the nucleus is much smaller than the inner radius of the disk, and that its mass is 37.8 million solar masses with a measurement uncertainty of only 0.3 percent (possible systematic errors due to the choice of model are larger, with about a 1 percent margin of error). This is by far the best case we have for a massive dark object at the center of any distant galaxy.[14]

Finally, our own galaxy offers unique evidence for a black hole.[15] Very close to the geometric center of the distribution of stars in the Milky Way is a compact source of strong radio emission known as Sagittarius A*. This region is difficult to study because small solid particles in interstellar space (commonly called "dust" but really more like haze or smoke) obscure visible light coming from stars near the center. The smoke can be penetrated by infrared radiation, and high-resolution observations at these wavelengths reveal a handful of bright stars within a few hundredths of a parsec from Sagittarius A*. The positions and velocities of these stars have been tracked, some for as long as two decades; in particular, the star S2 has an orbital period of only 15.8 years and now has been tracked through more than one complete orbit. Some four centuries ago, Johannes Kepler showed that the orbits of the planets around the sun were ellipses; here the orbit of S2 is also an ellipse (Figure 2). Using first-year mechanics, we can deduce from this orbit that the star is orbiting a body that is located at the radio source Sagittarius A*, that this body has a mass of 4.3 million solar masses, with an uncertainty of less than 10 percent, and that the size of this body is less than only one hundred times the Earth-sun distance, or a few thousand times the radius of the event horizon for a black hole of this mass. This extreme concentration of mass is incompatible with any known long-lived astrophysical system other than a black hole. The center of our galaxy thus offers the single best case for the existence of black

Figure 2
Orbits of Stars Near the Center of Our Galaxy



The radio source Sagittarius A*, believed to coincide with the black hole at the galaxy center, is at the zero point of the coordinates. The width of the frame is 0.03 parsecs or 6,700 times the Earth-sun distance. Each orbit is well-fit by an ellipse with one focus at Sagittarius A* (the focus is not on the symmetry axis of the orbit because the normal to the orbit is inclined to the line of sight). The smallest orbit, called S2, has a period of 15.8 years, and its point of closest approach to Sagittarius A* is 120 times the Earth-sun distance. These parameters imply that Sagittarius A* is associated with a mass of 4.3 million solar masses contained within about 100 times the Earth-sun distance. Source: Reinhard Genzel, Frank Eisenhauer, and Stefan Gillessen, "The Galactic Center Massive Black Hole and Nuclear Star Cluster," *Reviews of Modern Physics* 82 (4) (2010): 3121–3195.

holes and strongly suggests that the massive dark objects found in the centers of other galaxies are also black holes.

What else have we learned from these discoveries? First, black holes seem to be present in most galaxies, except perhaps for a class known as late-type galaxies. Second, the mass of the black hole is strongly correlated with the mass or luminosity of the galaxy; roughly, the black-hole mass is about 0.2 percent of the mass of the stars in the galaxy. But are the black holes we are finding in nearby galaxies really dead quasars? From galaxy surveys we can determine the average mass density in stars in the local universe, and since black-hole masses are typically 0.2 percent of the stellar mass in a galaxy, we can estimate the mass density of black holes in the local universe. Sołtan's argument, described earlier, gives the mass density of dead quasars in the local universe from completely different data (surveys of distant quasars as opposed to surveys of nearby galaxies). The two estimates agree to within a factor of about two – well within the uncertainties – so there is little doubt that the black holes we have found are indeed the ash from quasars or other active galactic nuclei.

This essay has described briefly what we have learned about the intimate relation between quasars, one of the most remarkable components of the extragalactic universe, and black holes, one of the most exotic predictions of twentieth-century theoretical physics. Many aspects of this relation remain poorly understood; to close, I will mention two of the most profound unanswered questions.

The first of these is the relation between black holes and galaxy formation. Although black holes make up only a fraction of a percent of the mass of the stars in galaxies, the energy released in forming them is hundreds of times larger than the energy released in forming the rest of the galaxy. If even a small fraction of the energy emitted by the black-hole furnace is fed back to the surrounding gas and stars, it would have a dramatic influence on the galaxy formation process. In an extreme case, the quasar feedback could blow the gas out of the galaxy and thereby quench the formation of new stars. Are black holes and quasars an interesting by-product of galaxy formation that has no influence on the formation process, or do they play a central role in regulating it? More succinctly, do galaxies determine the properties of quasars or vice versa?

The second profound question is one of physics rather than astronomy. All of the tests of Einstein's theory so far – which it has passed with flying colors – have been conducted in weak gravitational fields, such as those on Earth or in the solar system. In contrast, we have no direct evidence that the theory works in strong gravitational fields. Many naturally occurring processes near black holes in galaxy centers – tidal disruption of stars, swallowing of stars, accretion disks, and even black-hole mergers – may potentially be measured with the next generation of astronomical observatories. Can we understand these processes well enough to test the unique predictions of general relativity for physics in strong gravitational fields, and will Einstein turn out to be right?

ENDNOTES

[1] Textbooks on general relativity include Bernard Schutz, *A First Course in General Relativity*, 2nd ed. (Cambridge: Cambridge University Press, 2009); James B. Hartle, *Gravity: An Introduction to Einstein's General Relativity* (San Francisco: Addison-Wesley, 2003); Sean Carroll, *Spacetime and Geometry* (San Francisco: Addison-Wesley, 2004); and Robert M. Wald, *General Relativity* (Chicago: University of Chicago Press, 1984). Monographs on black holes include Subrahmanyan Chandrasekhar, *The Mathematical Theory of Black Holes* (Oxford: Clarendon Press, 1992); and Valeri P. Frolov and Andrei Zelnikov, *Introduction to Black Hole Physics* (Oxford: Clarendon Press, 2011). A popular account closely related to the subject of this essay is Mitchell Begelman and Martin J. Rees, *Gravity's Fatal Attraction*, 2nd ed. (Cambridge: Cambridge University Press, 2009).

[2] Subrahmanyan Chandrasekhar, *The Mathematical Theory of Black Holes* (Oxford: Clarendon Press, 1992), prologue.

[3] The physics of accretion disks is described in Marek A. Abramowicz and P. Chris Fragile, "Foundations of Black Hole Accretion Disk Theory," *Living Reviews in Relativity* 16 (2013): 1, arXiv:1104.5499; H. C. Spruit, "Accretion Disks" (May 2010), arXiv:1005.5279 (the essay is an expanded and revised version of "Accretion Theory," a lecture Spruit delivered at the XXI

Canary Islands Winter School of Astrophysics, Puerto de la Cruz, Tenerife, Spain, November 2 – 13, 2009); and Juhan Frank, Andrew King, and Derek Raine, *Accretion Power in Astrophysics*, 3rd ed. (Cambridge: Cambridge University Press, 2002). For a review of magnetohydrodynamic turbulence, energy dissipation, and angular momentum transport in accretion disks, see Steven A. Balbus and John F. Hawley, "Instability, Turbulence, and Enhanced Transport in Accretion Disks," *Reviews of Modern Physics* 70 (1998): 1 – 53.

[4] Introductory texts on galaxies include L. S. Sparke and J. S. Gallagher III, *Galaxies in the Universe: An Introduction*, 2nd ed. (Cambridge: Cambridge University Press, 2007); and Peter Schneider, *Extragalactic Astronomy and Cosmology* (Berlin: Springer, 2006). At a more advanced level there is James Binney and Michael Merrifield, *Galactic Astronomy* (Princeton, N.J.: Princeton University Press, 1998); James Binney and Scott Tremaine, *Galactic Dynamics*, 2nd ed. (Princeton, N.J.: Princeton University Press, 2008); and Houjun Mo, Frank van den Bosch, and Simon White, *Galaxy Formation and Evolution* (Cambridge: Cambridge University Press, 2010).

[5] Julian H. Krolik, *Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment* (Princeton, N.J.: Princeton University Press, 1999).

[6] For the history of this discovery, see K. I. Kellermann, "The Discovery of Quasars," *Bulletin of the Astronomical Society of India* 41 (2013): 1 – 17. An interview with Maarten Schmidt is available at http://oralhistories.library.caltech.edu/118/. A snapshot of the intense early debates about the nature of quasars can be found in George B. Field, Halton Arp, and John N. Bahcall, *The Redshift Controversy* (Reading, Mass.: W. A. Benjamin, 1973).

[7] Most measurements constrain the size of the so-called broad-line region, in which the broad optical emission lines of the quasar are produced. In black-hole models of quasars the broad-line region is much larger than the event horizon or accretion disk. Reverberation mapping is based on the average time delay between variations in the continuum luminosity, believed to arise from the accretion disk, and variations in the broad lines, which measure the light-travel time to the broad-line region. Photoionization models are based on an empirical correlation $R \propto L^{0.5}$ between the size of the broad-line region R revealed by reverberation mapping and the continuum luminosity L; this correlation is natural if the broad lines are in ionization equilibrium and brighter active galactic nuclei are simply scaled-up versions of fainter ones. For a review and references, see Yue Shen, "The Mass of Quasars," *Bulletin of the Astronomical Society of India* 41 (2013): 61 – 115. An alternative is to study quasars that are gravitationally lensed by an intervening galaxy; lensing by individual stars in the lens galaxy then leads to fluctuations in the brightness of the quasar image that depend on the ratio of the size of the broad-line region to the Einstein radius of the star. See E. Guerras et al., "Microlensing of Quasar Broad Emission Lines: Constraints on Broad Line Region Size," *The Astrophysical Journal* 764 (2013): 160.

[8] J. C. Miller, "Relativistic X-Ray Lines from the Inner Accretion Disks Around Black Holes," *Annual Review of Astronomy and Astrophysics* 45 (2007): 441 – 479.

[9] M. Boettcher, D. E. Harris, and H. Krawczynski, eds., *Relativistic Jets from Active Galactic Nuclei* (Weinheim, Germany: Wiley-VCH, 2012).

[10] Ronald A. Remillard and Jeffrey E. McClintock, "X-Ray Properties of Black-Hole Binaries," *Annual Review of Astronomy and Astrophysics* 44 (2006): 49 – 92.

[11] Andrzej Sołtan, "Masses of Quasars," *Monthly Notices of the Royal Astronomical Society* 200 (1982): 115 – 122.

[12] This process, known as dynamical friction, is a manifestation of energy equipartition familiar from statistical mechanics. See James Binney and Scott Tremaine, *Galactic Dynamics* (Princeton, N.J.: Princeton University Press, 2008).

[13] John Kormendy and Luis C. Ho, "Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies," *Annual Review of Astronomy and Astrophysics* 51 (2013): 511 – 653; Kayhan Gültekin et al., "The $M-\sigma$ and $M-L$ Relations in Galactic Bulges, and Determinations of Their Intrinsic Scatter," *The Astrophysical Journal* 698 (2008): 198 – 221; and Nicholas J. McConnell and

Chung-Pei Ma, "Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties," *The Astrophysical Journal* 764 (2013): 184.

[14] J. M. Moran, "The Black-Hole Accretion Disk in NGC 4258: One of Nature's Most Beautiful Dynamical Systems," in *Frontiers of Astrophysics: A Celebration of NRAO's 50th Anniversary*, ed. A. H. Bridle, J. J. Condon, and G. C. Hunt (San Francisco: Astronomical Society of the Pacific Conference Series, 2008): 395, 87; R. Herrnstein et al., "The Geometry of and Mass Accretion Rate Through the Maser Accretion Disk in NGC 4258," *The Astrophysical Journal* 629 (2005): 719 – 738; and Jenny E. Greene et al., "Precise Black Hole Masses from Megamaser Disks: Black Hole-Bulge Relations at Low Mass," *The Astrophysical Journal* 721 (2010): 26 – 45.

[15] For a comprehensive review of the central parsec of our galaxy, see Reinhard Genzel, Frank Eisenhauer, and Stefan Gillessen, "The Galactic Center Massive Black Hole and Nuclear Star Cluster," *Reviews of Modern Physics* 82 (4) (2010): 3121 – 3195.

The Formation & Evolution of Galaxies

Pieter van Dokkum

Abstract: Weighing in at 10^{42} kilograms and measuring 10^{21} meters across, galaxies are perhaps the most awe-inspiring objects known to mankind. They are also the only places in an otherwise dark and unforgiving universe where stars and planets are able to form. In the past five to ten years we have made enormous progress in understanding when galaxies came into being and how they changed and evolved over the course of cosmic time. For the first time, we have a rudimentary idea of what our own Milky Way looked like in the distant past, and we can now simulate Milky Way–like galaxies inside powerful computers. As we are starting to understand what happened in our galaxy's past, we are now turning to the question of why it happened. Untangling the complex physical processes that shape galaxies is extremely difficult, and will require continued advances in computers and information from powerful new telescopes coming online in the next decade.

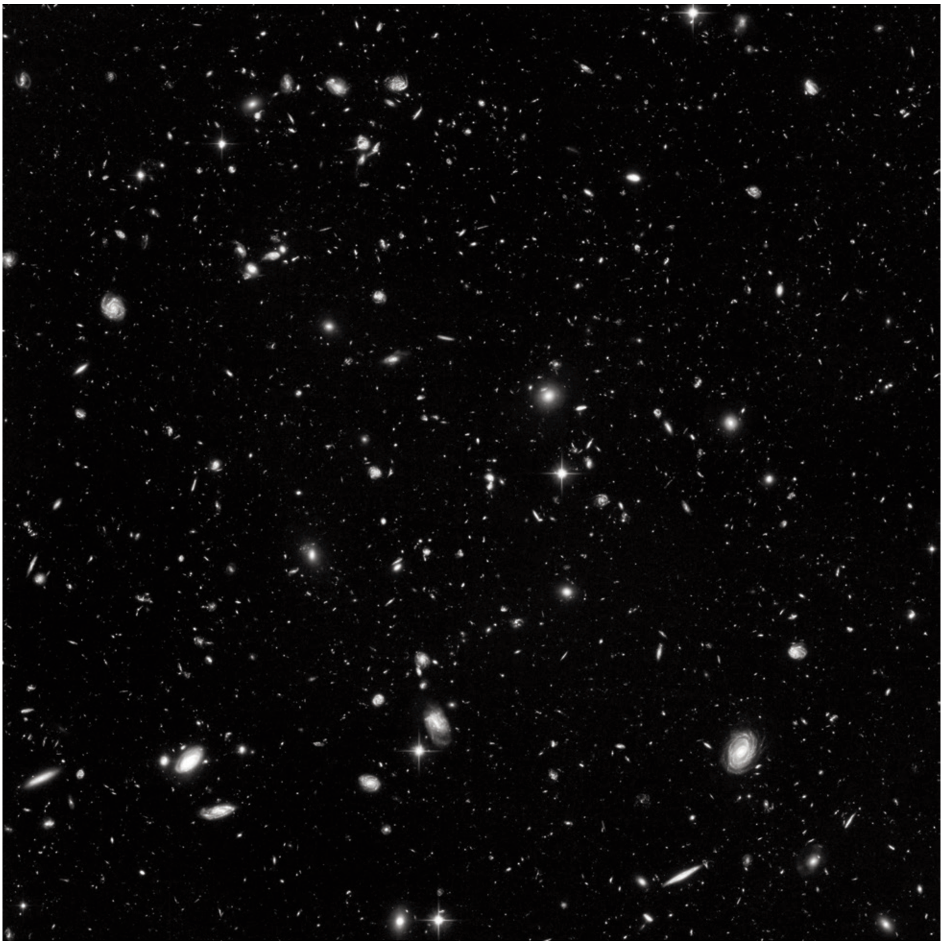
PIETER VAN DOKKUM is the Chair of the Department of Astronomy at Yale University, and the Sol Goldman Professor of Astronomy and of Physics. His research focuses on observational studies of the formation and evolution of galaxies. His many publications include articles in such journals as *Nature*, *The Astrophysical Journal*, and *The Astronomical Journal*.

If a forest is a collection of trees, and a city a collection of buildings, then the universe is a collection of galaxies. This is apparent when we look at the Ultra Deep Field, a remarkable image obtained with the Hubble Space Telescope that shows the faintest light humanity has yet detected (see Figure 1).[1] The blackness of space is punctuated by little blobs of light, each comprising a gravitationally bound system of tens of billions of stars. Galaxies contain nearly all the stars and planets in the universe, play host to the supermassive black holes in their centers, and serve as signposts delineating the large-scale cosmic web of dark matter structure. How galaxies were formed is a central question in astronomy. And because galaxies live at the intersection of the study of the structure of the universe as a whole and of the properties of the dark matter, gas, stars, and planets within them, the question is interwoven with many other fields of astronomy. Furthermore, understanding galaxy formation also means understanding our own galaxy, the Milky Way, and therefore our own cosmic history.

This is a young field of research. One hundred years ago, astronomers were trying to measure the extent

doi:10.1162/DAED_a_00311

Figure 1
The Hubble Ultra Deep Field



The Hubble Ultra Deep Field is a patch of sky observed for many days with the Hubble Space Telescope. Every little white blob is a distant galaxy, typically containing tens of billions of stars. Although the image covers only a tiny part of the sky (more than ten million of such patches would be needed to cover the entire sky!) it contains thousands of distant galaxies. Source: National Aeronautics and Space Administration and ESA/Hubble, http://www.spacetelescope.org/.

of the Milky Way, and the question whether the Milky Way makes up the entire universe or other galaxies exist beyond its boundaries was hotly debated. The cosmological framework for understanding galaxy formation and evolution was developed in the 1980s and 1990s,[2] and the first large surveys of the distant universe were undertaken in the early 2000s. There has been tremendous progress over the past decade, with some of the major questions that astronomers were struggling with now settled. We now have an idea of how galaxies came into being and how they changed over time. It is an incomplete story, with glaring omissions, inconsistencies, and unanswered questions; but a story nonetheless.

Largely thanks to surveys such as the Sloan Digital Sky Survey (see Michael Strauss's article in this volume for more on the Sloan survey) we now have a fairly complete census of galaxies in the "nearby universe": a loosely defined sphere with a volume of about a billion cubic light-years centered on our own galaxy. The Sloan survey has mapped about a million galaxies in this sphere, and we can study their luminosities, colors, morphologies, star formation rates, as well as other properties.

Luminous galaxies show a remarkable regularity in the nearby universe and can usefully be divided into two basic types. Most stars live in large *spiral galaxies* (as shown in Figure 2), which are characterized by majestic rotating disks of young stars with a dense central bulge of old stars. Our sun resides in such a galaxy: a piece of knowledge that reinforces the notion that "we are not in a special location," as Copernicus first put forth in 1543. Spiral galaxies continuously form new stars in their arms. The hot, short-lived massive stars that are formed in this process give the disks their characteristic blue color.

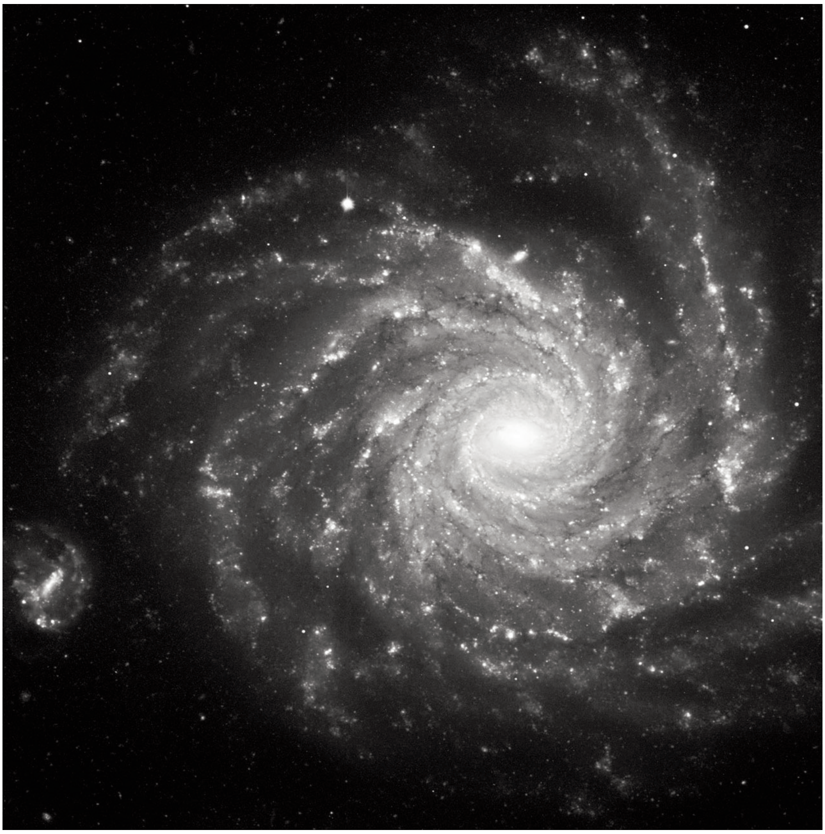
The other basic type of galaxy lacks spiral arms, is red, and has stopped forming new stars long ago (see Figure 3). Historically, these galaxies are called *early-type galaxies* since it was once thought that they represent an early stage of galactic evolution; and as often is the case, the name stuck though the interpretation evolved. One of the major results of the Sloan Digital Sky Survey is that these two basic galaxy types are well separated in many projections of the parameters of galaxy mass, kinematics, stellar age, color, luminosity, and galactic environment.[3] A galaxy is usually clearly either a spiral galaxy or an early-type galaxy: intermediate types are rare.

Significantly, it has also become clear that all galaxies are embedded in large structures composed of dark matter, somewhat confusingly termed "halos." The mass of a galaxy's dark matter structure is typically five to ten times greater than the mass of the rest of the galaxy, which means that it is the dark matter that dictates the processes driven by gravity. In many ways, dark matter controls galaxy evolution and the stars are just along for the ride. The nature of dark matter is still elusive because we have not yet identified a dark matter particle (if one even exists). Nevertheless, although we do not yet know what dark matter is, we have a very good idea of where it is. In fact, again using the Sloan survey, we can now statistically tie galaxies to their dark matter halos and offer a full description of the relation between dark matter and normal matter.[4] This description works remarkably well, and it has been tested by studying the gravitational lensing of faint background galaxies by the dark matter. It turns out that galaxies are also very regular in terms of the relationship between dark and conventional matter: when a galaxy's mass in normal matter is known, the mass in dark matter can be predicted with an accuracy of about 40 percent.

This regularity in galactic properties makes our task easier: we do not have to decipher the formation history of every individual galaxy. But we do have to understand how the two basic galaxy types came into being and why they are so distinct from one another. And once we understand how typical spiral galaxies were formed, we can apply this lesson to the Milky Way and learn about our own cosmic past.

Our galaxy and the other galaxies in the local volume of the universe hold important clues to their history. The great majority of stars that have formed in the history of the universe are still around today, and the present-day appearance of a galaxy is the accumulation of all the things that happened to it over the course of cos-

Figure 2
A Spiral Galaxy



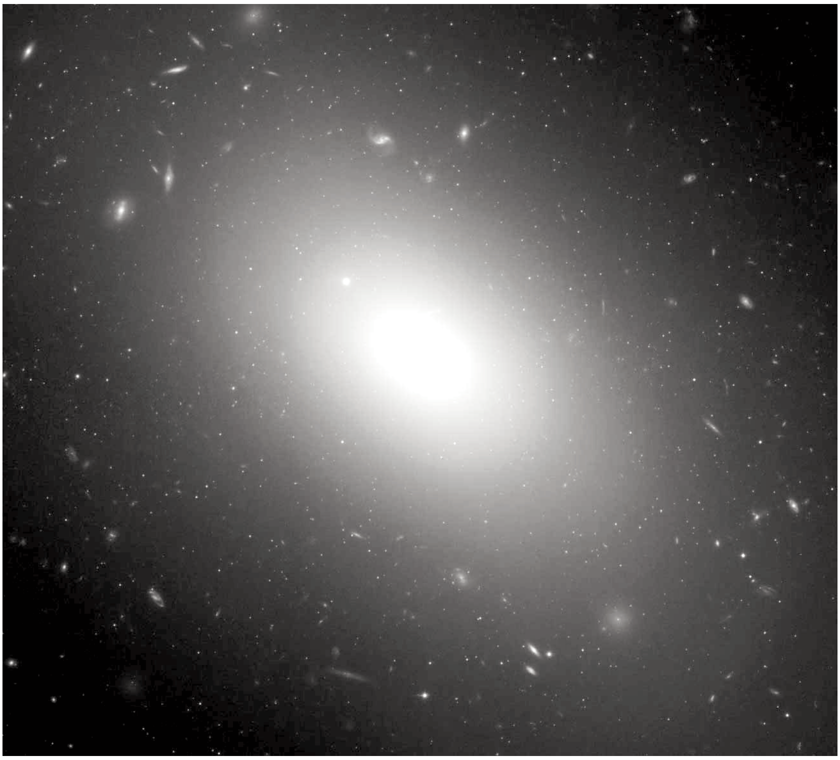
Galaxies come in two basic types. This is a spiral galaxy, characterized by a spinning disk with its eponymous spiral arms. In these spiral arms hydrogen gas is continuously processed into new stars. The Milky Way is a spiral galaxy; the Orion nebula is a region of intense star formation in a spiral arm relatively close to us. Source: The Dragonfly Telephoto Array, http://dunlap.utoronto.ca/instrumentation/dragonfly/; Roberto Abraham and Pieter van Dokkum.

mic time. Just like a paleontologist reconstructs a dinosaur from bone fragments, we can use this "fossil evidence" of earlier epochs to reconstruct what galaxies looked like in the past.

From studies of the Milky Way and other nearby galaxies, it appears that the universe is currently in a much more sedate state than it was in the past. Most galaxies form new stars at a relatively low pace, and so to have built up the vast reservoirs of existing stars that we observe today their formation rates must have been much higher in the past. Studies of the ages of the various components of the Milky Way and its neighbors suggest that the distinction between spiral galaxies and early-type galaxies, which is so characteristic of the distribution of galaxies, may be a transient phenomenon in the history of the universe. Spiral disks are relatively young and may have arrived on the scene in their present form only in the past five to eight billion years.[5]

Perhaps the most spectacular result of studies of nearby galaxies lies hidden in

Figure 3
Early-Type Galaxies



Basically smooth big balls of stars, early-type galaxies constitute the other basic type of galaxy. In these galaxies all the stars are old, and new star formation has not occurred for many billions of years. Source: National Aeronautics and Space Administration and ESA/Hubble, http://www.spacetelescope.org/.

their outskirts. Sensitive stellar mapping programs of the Milky Way and its neighbor M31 (the Andromeda galaxy) have demonstrated that they are embedded in a vast network of stellar streams, the debris of previous encounters with other galaxies.[6] Such features had first been seen around some galaxies many decades ago, but it is now thought that all galaxies may have vast debris fields around them, although often just below the detection threshold of present-day instrumentation. These streams point toward a violent past when interactions and collisions among galaxies were much more common than they are today.

Galactic paleontology has fundamental limitations, just as looking at dinosaur bones only gives us incomplete and fragmented information on living, breathing dinosaurs. Galactic collisions and mergers erase much of the past history of a galaxy, making it difficult to discern how it was built up. Furthermore, other processes such as bar instabilities and stellar migration gradually change the appearance of galaxies over time even if they do not experience collisions and are instead left to their own devices. As a result, the key building phases of today's galaxies cannot be deciphered from their present-day appearance alone. Luckily, we are not limited to our own neighborhood, and we can do something that dinosaur-hunting paleontologists can only dream of.

Owing to the finite speed of light, we can directly observe the past. Looking out into space, we see the moon as it was a second ago, the sun as it was eight minutes ago, and the Andromeda galaxy (the most distant object visible to the unaided eye) as it was 2.5 million years ago. Two and a half million years is only 0.17 percent of the 14.8-billion-year-old universe, which is why we (again, somewhat loosely) consider the local volume representative of the present-day universe.

Using large telescopes on Earth and in space we are able to detect and study galaxies well beyond the local volume, at distances where the look-back time is a significant fraction of the age of the universe. Until a few decades ago, we could identify samples of galaxies at distances of about seven billion light-years, allowing us to look back in time about half the age of the universe. In the mid-1990s, with the combination of Hubble in space and the Keck telescopes on Earth, we learned to take sharp images of galaxies over 95 percent of cosmic time, lifting a veil from the early universe. This frontier is continuously pushed farther into the past, as new detector technology greatly expands the capabilities of existing telescopes. The final space shuttle servicing mission of the Hubble Space Telescope was of particular importance, as the new instruments improved the sensitivity of Hubble by factors of five to twenty.

One of the new cameras installed on the Hubble during that 2009 servicing mission has been used over the past three years to capture images for one of the largest-ever projects undertaken by the telescope, which aims to determine how galaxies were assembled. Using more than one thousand two hundred hours of observing time, the CANDELS, 3D-HST, and HUDF09+12 projects are providing us with samples of hundreds of thousands of galaxies at cosmological distances, with well-measured sizes, star formation rates, masses, and colors. These data allow us to create snapshots of the universe at different epochs and to study how the distribution of galaxy properties has changed over time. This measurement is essentially statistical in nature: we cannot see the whole of the Milky Way when we look out into space, but we can see galaxies that are like the Milky Way. We are also learning how to "connect the dots"; that is, to find which galaxy populations at early epochs were ancestors of which galaxy populations today.

The Hubble observations, aided by studies at other wavelengths and with large telescopes on Earth, paint a picture of dramatic change. Ten billion years ago galaxies were two to four times smaller than they are today, and yet the rate of star formation was ten times greater. Combining these results, the density of star formation (how many new stars are formed in a fixed region of space within a galaxy) was up to one hundred times greater in these early-universe galaxies than it is in galaxies today.[7] Not surprisingly, the early-universe galaxies also look very different from galaxies today: they are bluer and have a more irregular appearance. The grand spiral galaxies with gently spinning thin disks that are now so ubiquitous were rare in the early universe.

The high star formation rates of early-universe galaxies tell us that they built up rapidly – so rapidly that many doubled their mass in less than a billion years. Eleven billion years ago, the Milky Way was a faint little blob with only 10 percent of its present-day stellar content but a very large amount of gas: the fuel for star formation. Over the next three to four billion years it proceeded to convert this gas into stars at a ferocious rate, adding about ten times the mass of the sun every year. It then gradually quieted down, became redder as its stellar population aged, and settled in its current spiral galaxy morphology about five billion years ago. Over this entire period, Milky Way–like galaxies increased their mass by a factor of ten and grew in size by a factor of two.[8]

Interestingly, some galaxies did not participate in the overall gas feeding frenzy in the young universe. Despite the availability of large amounts of fuel, about 50 percent of the most massive galaxies stopped forming new stars as early as ten billion years ago. This strange reluctance to form new stars was already suggested by the old ages of stars in present-day massive galaxies, and it has now been confirmed by direct observations of massive "dead" ancestor galaxies in the young universe made with the Hubble, Keck, and Gemini telescopes.[9] The structure of these ancestors, as revealed by the Hubble Space Telescope, did yield a surprise. It turns out these galaxies were extremely compact in the past, much smaller than they are today. Remarkably, their sizes increased by a factor of four over the past ten billion years, whereas their masses increased by only a factor of two.

Overall, the epoch around eight to ten billion years ago was characterized by a high degree of diversity. We see very compact "dead" galaxies, large and thick star-forming disks, dust-enshrouded collisions, and many other galaxy types. This was also the era when massive black holes at the centers of galaxies were growing rapidly: many galaxies show activity in their nuclei that cannot be explained by star formation and instead reveals the energetic processes associated with black hole growth. This period has been described as "high noon," the "heyday of galaxy formation," or – with a nod to our colleagues in the biological sciences – the "cosmic Cambrian," as the universe seemed to be experimenting with galaxy shapes and sizes.

Our views of even earlier epochs are necessarily less complete, since the feeble light of the galaxies comes from even greater distances. Nevertheless, using the deepest images of the night sky ever obtained (see again Figure 1) astronomers have identified galaxies to within a few hundred million years of the Big Bang and characterized their star formation rates, sizes, and other properties.[10] It now appears that the average star formation rate in the universe increased rapidly in the first billion years after the Big Bang, after which it had a broad peak and then declined. Interestingly, we can now study galaxies at epochs when the hydrogen that exists in the vast spaces between them was still partially neutral; it is an open question whether the ultraviolet radiation of the hot, massive stars in these early galaxies were responsible for ionizing the universe.

As observations of the universe are providing an increasingly detailed description of the properties of galaxies, the research focus is not only on what happened but also on why it happened. There is broad consensus on the general outline of the process of galaxy formation: gravity and the expansion of the universe rule the behavior of dark matter, and give rise to dark matter objects with a distribution of masses that roughly follows a power law (wherein the mass of dark matter can be predicted approximately as a power of the mass of conventional matter). Gas initially follows the distribution of dark matter but then cools and forms stars. The efficiency of this process depends on the dark matter mass, such that the final distribution of the stellar masses of galaxies is not a power law, but has a preferred scale around the mass of the Milky Way.

The details of these processes are fiendishly complex because the dark matter, gas, and stars are all intertwined and continuously changing as structures grow. Futhermore, the relevant physical processes happen on an enormous range of scales, from the distances between galaxies to the central regions of the birth clouds of individual stars: a range that spans thirteen orders of magnitude. To put this challenge into context, it is equivalent to simultaneously understanding the processes that operate on the scale of the Earth-moon system and processes that operate on the scale of the width of a human hair. As if this were not difficult enough, the relevant time scales range from the billions of years of galaxy interactions to the ten-thousand-year freefall timescale of protostellar clouds.

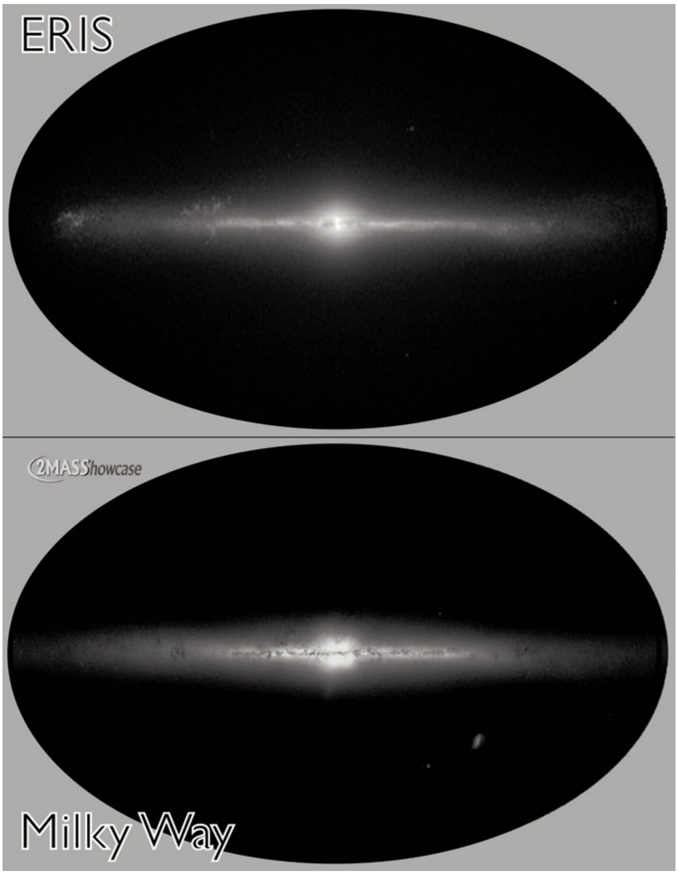
Despite these seemingly insurmountable challenges, in the past five years there have been some remarkable successes in modeling the process of galaxy formation. These have mostly come from advanced algorithms running on the world's fastest computers, a method that has its roots in the first galaxy formation simulations done in the late 1960s. The simulations treat the dark matter and stars as collisionless particles, where a single "particle" usually stands for some ten thousand actual stars. Gas is treated with hydro-dynamical techniques, which simulate the flow of gas and incorporate cooling and heating. The problem of the vast range of scales is ameliorated by adaptively changing the physical scale in the simulation, using a coarse grid for the space in between galaxies and a very fine grid inside the star-forming complexes of spiral galaxies. As even this fine grid cannot capture the formation of individual stars, analytical prescriptions are used for the "subgrid physics"; that is, the processes that happen on scales that are not resolved by the simulation.

The simulations provide us with movie clips showing the formation of galaxies over the entire history of the universe, sped up by a factor of 10^{16}. They show how little star-forming fragments assemble from gas clouds in the first billion years after the Big Bang, how these fragments grow and merge with one another, how spinning disks form from gas that either condenses gradually out of the dark matter halo or is injected by cold streams, how these disks are destroyed or puffed up in subsequent collisions with other galaxies, and how massive galaxies continue to grow by consuming their little neighbors.

Several results stand out. Perhaps the most impressive is that we now have artificial galaxies living in computers that look a lot like normal spiral galaxies and early-type galaxies in the actual universe (Figure 4).[11] This is an outstanding achievement: until recently it was not possible to start a simulation shortly after the Big Bang and end up with anything that remotely resembled the Milky Way. In terms of the relevant physical processes, a critical breakthrough was the discovery that gas can get into the central regions of dark matter halos via two paths: it can be shock-heated to the virial temperature of the halo followed by gradual cooling, and it can flow directly to the center along a filament or stream.[12] The simulations have also demonstrated the importance of mergers with small galaxies for the growth of massive early-type galaxies: the cores of massive galaxies form first, and then their outer envelopes are added gradually by accretion.[13] Finally, the simulations consistently demonstrate the overwhelming importance of feedback processes; that is, how much energy is returned to the interstellar medium by newly formed stars and black holes.

These accomplishments come with several crucial asterisks. Many of the key processes, in particular those relating to stellar and black hole feedback, take place on unresolved scales (the subgrid), which means they are essentially free parameters in the simulations. Furthermore, galaxies

Figure 4
Computer Simulation of the Milky Way Compared to the Actual Milky Way



In the past few years, astronomers have managed to create artificial galaxies that look very much like real galaxies. Here, the computer simulated Milky Way sits atop a photographic image of the Milky Way. Source: Javiera Guedes, Simone Callegari, Piero Madau, and Lucio Mayer, "Forming Realistic Late-Type Spirals in a ΛCDM Universe: The Eris Simulation," *The Astrophysical Journal* 742 (76) (2011), arXiv:1103.6030.

in our computer simulations simply love forming stars: they are much more efficient at it than the real ones appear to be. We can artificially lower the star formation efficiency by tuning the subgrid parameters, but this is a far cry from understanding the physical processes involved. A possibly related issue is that it has proven difficult to match all observations at once. For example, the simulations that successfully produce early-type galaxies have difficulty producing realistic Milky Way–like galaxies. Finally, the simulations are typically tuned to match the appearance of galaxies in the present-day universe, and they may not fare as well when compared to the newly available observational data on galaxies at earlier times.

Over the past decade – and even in the last five years – we have made dramatic progress in our understanding of galaxy formation and evolution. We now have a broad idea of the processes that governed the fourteen-billion-year-long home improvement project that was the making of the Milky Way. At the same time, we are only at the beginning: we now have a better idea of the problems we need to solve (such as the too-high star formation efficiency in model galaxies); but that is quite different from actually having the solutions. Additionally, our information on distant galaxies – while greatly improved over the course of the past decade – is still crude by most standards, since we typically only have a few characteristic numbers to work with: galaxies' luminosities, colors, sizes, and a rough measure of their star formation rates. Perhaps most fundamentally, this review did not touch on the biggest questions of all: until we pin down the nature of dark energy and dark matter, we can hardly claim to understand the formation of structure in the universe.

As new observing facilities come online and computers and computer algorithms continue to improve in the next decade, and as we gain a better understanding of the physical processes driving galaxy formation through advances in other subfields of astronomy, we can expect further progress. The ALMA facility in Chile will provide us with detailed images of distant galaxies in the light of molecular gas, allowing us to directly connect the existing stars to the fuel for new star formation. The James Web Space Telescope, the successor to the Hubble Space Telescope, will open up the earliest epochs of galaxy formation for study and provide vastly deeper and higher resolution views of distant galaxies than we have access to now. New ground-based telescopes will explore both the large scale distribution of galaxies and provide high resolution images and spectroscopy of small samples. Finally, the GAIA mission will provide a three-dimensional map of about a billion stars in our own Milky Way, allowing us to piece together our own history via "galactic paleontology" on a vast scale.

ENDNOTES

[1] Steven V. W. Beckwith et al., "The Hubble Ultra Deep Field," *The Astronomical Journal* 132 (2006): 1729.

[2] Simon D. M. White and Carlos S. Frenk, "Galaxy Formation through Hierarchical Clustering," *The Astrophysical Journal* 379 (1991): 52.

[3] Guinevere Kauffmann et al., "Stellar Masses and Star Formation Histories for 10^5 Galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society* 341 (2003): 33.

[4] Charlie Conroy and Risa H. Wechsler, "Connecting Galaxies, Halos, and Star Formation Rates across Cosmic Time," *The Astrophysical Journal* 696 (2009): 620.

[5] Hans-Walter Rix and Jo Bovy, "The Milky Way's Stellar Disk," *The Astronomy and Astrophysics Review* 21 (1) (2013), arXiv:1301.3168.

[6] Alan W. McConnachie et al., "The Remnants of Galaxy Formation from a Panoramic Survey of the Region around M31," *Nature* 461 (2009): 66.

[7] L. J. Tacconi et al., "High Molecular Gas Fractions in Normal Massive Star-Forming Galaxies in the Young Universe," *Nature* 463 (2010): 781.

[8] Pieter G. van Dokkum et al., "The Assembly of Milky Way–Like Galaxies since z~2.5," *The Astrophysical Journal Letters* 771 (2013), arXiv:1304.2391.

[9] Mariska Kriek et al., "Spectroscopic Identification of Massive Galaxies at z~2.3 with Strongly Suppressed Star Formation," *The Astrophysical Journal* 649 (2006): L71.

[10] R. J. Bouwens et al., "Ultraviolet Luminosity Functions from 132 z~7 and z~8 Lyman-Break Galaxies in the Ultra-Deep HUDF-09 and Wide-Area Early Release Science WFC3/IR Observations," *The Astrophysical Journal* 737 (2011): 90.

[11] Javiera Guedes, Simone Callegari, Piero Madau, and Lucio Mayer, "Forming Realistic Late-Type Spirals in a Lambda-CDM Universe: The Eris Simulation," *The Astrophysical Journal* 742 (2011): 76; and L. Oser, T. Naab, J. P. Ostriker, and P. H. Johansson, "The Cosmological Size and Velocity Dispersion Evolution of Massive Early-Type Galaxies," *The Astrophysical Journal* 744 (2012): 63.

[12] D. Keres, N. Katz, D. H. Weinberg, and R. Dave, "How Do Galaxies Get Their Gas?" *Monthly Notices of the Royal Astronomical Society* 363 (2005): 2.

[13] Oser et al., "The Cosmological Size and Velocity Dispersion Evolution of Massive Early-Type Galaxies."

Cosmology Today

David N. Spergel

Abstract: We seem to live in a simple but strange universe. Our basic cosmological model fits a host of astronomical observations with only five basic parameters: the age of the universe, the density of atoms, the density of matter, the initial "lumpiness" of the universe, and a parameter that describes whether this lumpiness is more pronounced on smaller physical scales. Our observations of the cosmic microwave background fluctuations determine these parameters with uncertainties of only 1 to 2 percent. The same model also provides an excellent fit to the large-scale clustering of galaxies and gas, the properties of galaxy clusters, observations of gravitational lensing, and supernova-based measurements of the Hubble relation. This model implies that we live in a strange universe: atoms make up only 4 percent of the visible universe, dark matter makes up 24 percent, and dark energy – energy associated with empty space – makes up 72 percent.

Cosmology is a historical science. Because light travels at a finite speed, when we look out in space, we look back in time. We see the sun as it was eight minutes ago, and we see nearby stars as they were five, ten, or a hundred years ago. It takes light approximately 2.5 million years to travel from the Andromeda galaxy to our eyes, so when we stare at our nearest major neighbor with a telescope, we observe Andromeda as it was back before the dawn of man. The farther out we look, the farther back we look in time. When the Hubble Space Telescope observes a distant galaxy, it sees the galaxy as it was perhaps 12 billion years ago. Our observations of the cosmic microwave background involve the oldest light, photons that formed only one year after the Big Bang and last interacted with atoms just four hundred thousand years after the Big Bang. This light travels for 13.7 billion years before reaching us, and it brings us our universe's baby picture.

Our basic model of cosmology rests on Einstein's nearly century-old theory of general relativity. As my late academic great grandfather Johnny Wheeler used to teach, "General Relativity consists of two simple ideas: matter tells space how to curve and space tells matter how to move." On the scale of our

DAVID N. SPERGEL, a Fellow of the American Academy since 2012, is the Charles A. Young Professor of Astronomy on the Class of 1897 Foundation and Professor of Astrophysical Sciences at Princeton University. He is a theoretical astrophysicist with interests ranging from the search for planets around nearby stars to the shape of the universe.

doi:10.1162/DAED_a_00312

solar system, the mass of the sun curves space around it, and our Earth moves on a nearly circular orbit in this curved space. On the cosmological scales, the distribution of matter is nearly uniform. General relativity implies that a nearly uniform universe must be either expanding or contracting. Since Edwin Hubble's observations in the 1920s, we have known that our universe is expanding.

As the universe expands, the distance between galaxies grows. Today, it takes light roughly fifty million years to travel to the Virgo cluster. Eight billion years ago, the distance between objects was a factor of two smaller, so light would have taken only twenty-five million years to travel from our galaxy to the Virgo cluster. As we go farther back in time, objects get closer and closer together. Thus, the early universe was much denser than today's universe.

This expansion of the universe not only increases the distance between galaxies, but also stretches light. Light emitted from a distant galaxy is "redshifted." If a galaxy eight billion years ago emits blue light, the radiation's wavelength is stretched as it travels toward us, and we observe the light as red. Because astronomers can easily measure the wavelength of light detected from a distant galaxy, we can determine its redshift and then use general relativity (and our cosmological model) to relate the redshift of a galaxy to its age. Today, our universe is 13.8 billion years old. When we observe a galaxy at redshift 1, the wavelength of the light has been stretched by a factor of two due to the expansion of the universe. It took light 8 billion years to travel from the galaxy to us, so we observe this distant galaxy as it was 5.7 billion years after the Big Bang. Yellow light emitted by the galaxy at redshift 1 at a wavelength of 550 nanometers will appear to us in the infrared at 1100 nanometers. The most distant known galaxy is at redshift 10. When optical light left this infant galaxy 13.3 billion years ago, it was only 500 million years after the Big Bang. Today, we see this light as infrared radiation and use our observations to study the properties of early galaxy formation.

General relativity relates the expansion rate of the universe to the density and geometry of the universe. If the energy in expansion exceeds the self-gravity of the matter in the universe, then the universe is negatively curved and will expand forever, growing increasingly cold and empty. On the other hand, if the energy in expansion is less than the self-gravity of the universe's matter, the expansion will slow down and reverse, and the universe will collapse in a future big crunch. As Robert Frost prophesized, the universe will end in either fire or ice.

Until recently, cosmologists thought that the expansion rate of the universe would slowly decelerate. The expansion rate of the universe is proportional to the square root of the density of the universe. Since the density of matter decreases as the universe expands, astronomers assumed that the expansion rate of the universe has been slowing with time.

But over the past thirty years, there has been growing evidence that the expansion rate of the universe has been increasing with time.[1] This result has shocked physics: the equivalent of throwing a ball upward and finding that gravity makes it accelerate away from the point of release. If general relativity is correct, this cosmic acceleration implies that most of the energy in the universe is in the form of dark energy: energy associated with empty space. In the late 1990s, measurements of the relationship between the distance and the redshift to supernova – powerful explosions of nearly uniform brightness that can be seen at very large distances – provided the strongest evidence for this strange phenomenon.[2] Soon afterward, measurements of the cosmic microwave background fluctuations confirmed this surprising cosmology.[3]

Dark energy is different from "dark matter." Ever since Fritz Zwicky's work in the 1930s, astronomers have suspected that stars are not the dominant form of matter in galaxies. By the 1970s, astronomers had assembled several independent lines of argument all implying that dark matter was neither gas nor stars. Dark matter appears to be some new type of particle that has not yet been found in our particle accelerators. Dark energy is even stranger: it does not cluster in galaxies, nor does it seem to respond to any of the natural forces. Dark energy affects the universe only through changing its expansion rate.

The cosmic microwave background radiation is the oldest light in the universe, the leftover heat from the Big Bang. This radiation fills all space and was once the dominant form of energy in the universe. The expansion of the universe cools the cosmic background radiation. Today, the temperature of the radiation is 2.73 degrees K. When the distance between galaxies was half its present value, the temperature of the cosmic background radiation was twice its present value. When the distance between galaxies was a tenth its present value, the temperature of the cosmic background radiation was ten times its present value.

As we go farther back in time and closer to the moment of the start of the Big Bang expansion, the universe is ever hotter. One second after the Big Bang, the temperature of the universe was 10 billion degrees C, and the universe was a nearly uniform sea of electrons, protons, neutrons, dark matter, and radiation. At that time, most of the energy density in the universe was in the form of radiation. Three minutes after the Big Bang, the temperature of the universe was about 500 million degrees C. During this period in the universe's evolution, most of the deuterium and helium in the universe was synthesized from neutrons and protons. Our measurements of the abundance of these two cosmic fossils are a direct determination of the density of the atoms at this early epoch.

During the first three hundred thousand years of cosmic history, almost all of the atoms in the universe were ionized into a plasma of electrons, protons, and helium ions. The cosmic background photons were frequently colliding with the electrons in this primordial plasma, so both atomic matter and photons were coupled together in a single fluid. As the universe cooled, the protons and helium ions were able to combine with electrons and form neutral hydrogen and helium atoms. By four hundred thousand years after the Big Bang, most of the electrons had combined with ions, and the universe was mostly neutral. Since the cosmic background photons do not interact with these neutral gases, they were able to propagate freely. The photons that we observe when we look at the cosmic background radiation last interacted with atoms at this very early time. Thus, when we observe the background radiation, we are directly measuring physical conditions at this early moment in the history of the universe.

In 1964, astronomers Arno Penzias and Robert Wilson detected the cosmic background radiation with their horn antenna at Bell Laboratories. Twenty-five years later, the COBE satellite found that this nearly uniform microwave radiation had exactly the spectral properties predicted by the hot Big Bang model. This measurement of the cosmic background is one of the foundational observations for the hot Big Bang model.

While the cosmic microwave background radiation is nearly uniform, there are tiny variations in the temperature of the radiation. These variations are primarily due to the fluctuations in the density and temperature of the universe four hundred thousand years after the Big Bang, the period when the electrons and protons combined to make hydrogen. COBE made the first detection of the micro-Kelvin-level variations in the microwave background temperature.[4] Subsequent observations by ground-based and balloon-based radio antennas mapped this background with ever improving technologies. In 2003, NASA's WMAP released its first detailed all-sky map of the fluctuations.[5] In 2013, ESA's Planck satellite team provided an even more detailed map that traces the same fluctuations.[6] Observations from WMAP, Planck, and ground-based telescopes such as the Atacama Cosmology Telescope in Chile and the South Pole Telescope in Antarctica give a remarkably consistent picture of the physical conditions in the early universe.[7]

The few millionth-of-a-degree temperature variations in the microwave background radiation seen by these experiments trace variations in the density and temperature of the early universe. Because the microwave background photons have been traveling to us with minimal interactions with intervening matter since four hundred thousand years after the Big Bang, these fluctuations reflect physical conditions at these early times.

Four hundred thousand years after the Big Bang, the early universe was a simple place. Electron, protons, and photons were bound together into a warm 3000 K plasma. Tiny variations in the density of the universe generated sound waves in this plasma. The distance that the sound waves could move in four hundred thousand years imparted a characteristic scale on the universe, and the self-gravity of the plasma and the dark matter determined the height of the peaks. Because these variations were small, cosmologists can use linear theory to accurately predict the relationship between the statistical properties of the fluctuations and the conditions in the early universe.

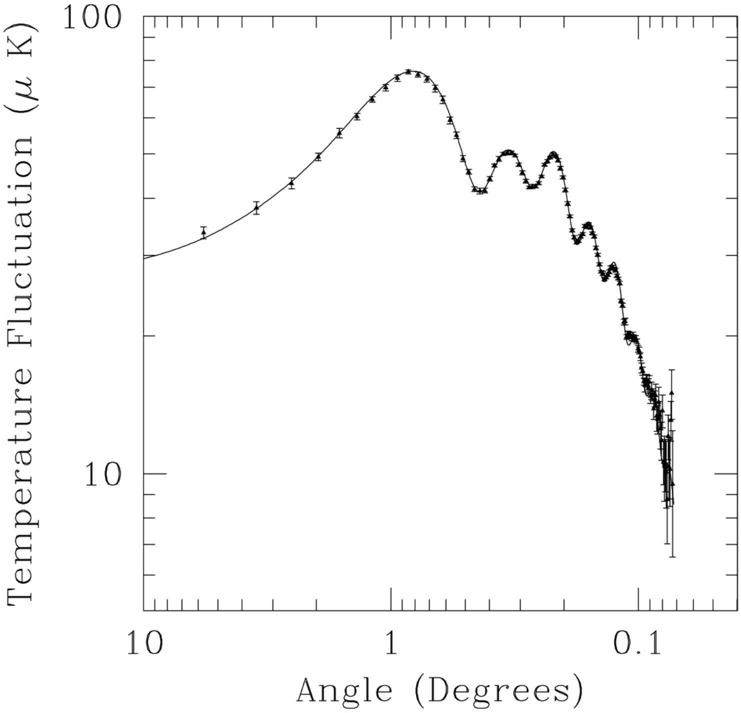
Cosmologists quantify the properties of these fluctuations by measuring their statistical properties. These fluctuations have very simple statistical properties: they are spatially homogenous and can be characterized almost entirely through measurements of the point correlation function of the data or, equivalently, the angular power spectrum. Figure 1 shows the measured angular power spectrum from the Planck satellite. The x-axis on this plot shows the angular size of the fluctuations; the y-axis shows the amplitude of the temperature fluctuations.

We can use the measurements of the amplitude of the peaks in the temperature angular power spectrum to infer the basic parameters of modern cosmology. The ratio of the height of the first peak to the second peak depends on the density of atoms in the early universe. The position of the peaks depends on the geometry of the universe. The height of the third peak is sensitive to the total density of matter. With our current observations of the cosmic microwave background, we can pin down all of the basic parameters to the precision of one part in a hundred.

We can use observations of the nearby universe to infer the same parameters through very different methods:

- Measurements of the abundance of deuterium and helium 4 provide determinations of the density of atoms accurate to 10 percent and consistent with the density inferred from the height of the peaks in the microwave background angular power spectrum.[8]
- Measurements[9] of the distances to nearby supernovae and Cepheid stars measure the expansion rate of the universe to be 74 km/s/Mpc, within 10 percent

Figure 1
Amplitude of Temperature Fluctuations as a Function of Angular Scale



The curve is the best-fit model based on the data from the Wilkinson Microwave Anisotropy Probe and the Atacama Cosmology Telescope; the points and their error bars are computed from a combination of the publicly available Planck satellite 217 GHz and 545 GHz data. Source: David Spergel, Raphael Flauger, and Renee Hlozek, "Planck Data Reconsidered," submitted to *Journal of Cosmology and Astroparticle Physics* (2013), arXiv:1312.3313.

of the value inferred from the cosmic microwave background observations. Cepheids are variable stars with a known relationship between their period and their luminosity. Because the intrinsic luminosity of supernovae and Cepheids are known, they are used as standard candles to measure distance. A mild discrepancy between these two measurements is a subject of active discussion at cosmology meetings.

- Supernova observations can also be used to trace the relationship between distance and expansion. These observations[10] also yield the cosmological parameters consistent with the cosmic microwave background observations and require a universe filled with dark matter, dark energy, and atoms. In 2011, the Nobel Prize for Physics committee recognized the leaders of these observations for their discovery of cosmic acceleration. In our simplest cosmological models, dark energy is the driver of this cosmic acceleration.
- Measurements[11] of the abundance of rich clusters of galaxies provide an alternative method of measuring the density of the matter and the amplitude of the primordial fluctuations. These mea-

surements again fit our basic cosmological model with a consistent set of parameters.

- The same sound waves that produce a characteristic scale in the microwave sky also produce a characteristic scale in galaxy clustering. Using the Sloan Digital Sky Survey, astronomers have now measured the positions of millions of galaxies. They can compute the statistical correlations of these galaxies and infer the density fluctuations in matter in the nearby universe.[12] These measurements agree remarkably well with the cosmic microwave background observations.

Despite the remarkable success of the Big Bang model in describing the evolution of the universe and the growth of fluctuations, the model is incomplete. Intriguingly, inflation – currently the most popular extension of the Big Bang model – not only addresses these problems but also makes predictions that we can test with our cosmic microwave background observations.[13]

The standard Big Bang model has a number of profound philosophical problems: it does not explain why our universe is so large, why the kinetic energy of our universe nearly perfectly balances the gravitational energy, or why the universe is nearly (but not perfectly) uniform. Our universe is more than 13.7 billion light years across, yet the "characteristic" scale set by general relativity and quantum mechanics is the Planck length, only 10^{-36} meters. Our nearly flat universe is at an unstable fixed point. Today, the kinetic energy of the universe (the energy in the expanding galaxies) is within 1 percent of the gravitational energy of the universe. Since these two quantities tend to rapidly evolve away from each other in the standard cosmology, they would have to be finely tuned to be nearly identical to more than twenty digits at the epoch of nucleosynthesis, the period of the universe three minutes after the Big Bang, when most of the deuterium and helium in the universe was synthesized.

The near, but not perfect, uniformity of the early universe is another puzzle in Big Bang cosmology. Different regions of space that were never in causal contact in the Big Bang model have nearly identical densities. The solution to the problem must explain this near, but not perfect, equality; for if the early universe were perfectly uniform, it would still be uniform today.

The inflationary paradigm offers an answer to these questions. It posits that the very early universe underwent an extremely rapid period of exponential expansion. Cosmologists call this very rapid period of expansion "inflation." This rapid expansion stretched the universe to a very large size. During this rapid period of expansion, the kinetic energy of the universe was driven to match the gravitational energy, and this enormous stretching erased any initial fluctuations in the early universe. When the inflationary paradigm was proposed thirty years ago, cosmologists recognized that the model not only solved these Big Bang cosmology problems, but also offered a mechanism to produce the fluctuations that would grow to form galaxies.

During the inflationary expansion, tiny quantum mechanical fluctuations in density were amplified enormously. Some regions of the universe had slightly higher densities while other regions of the universe had slightly lower densities. The regions with slightly higher densities spent more time in the exponential expansion phase and exited inflation later. After the universe cooled and became dominated by matter, these denser regions grew and eventually collapsed to form galaxies, stars, and planets. Thus, the inflationary model implies that the origin of all of the structure in the universe was the tiny quantum mechanical fluctuations amplified during the first moments of the Big Bang.

This remarkable explanation for the origin of structure is testable. The inflationary model is highly predictive about the statistical properties of these fluctuations. This instability during the inflationary phase led to a very specific prediction for the statistical properties of the variations in density: the fluctuations should be "Gaussian random phase, adiabatic, nearly scale-invariant" fluctuations. Gaussian random phase fluctuations have very simple statistical properties and are described entirely by their two-point correlation function. Adiabatic fluctuations have the same ratio of photons, electrons and protons, and dark matter everywhere. Scale-invariant fluctuations have the same amplitude on all scales. Thus, the inflationary model predicts that the statistical properties of the temperature of the tens of millions of points in the Planck satellite maps and the statistical properties of the positions of millions of galaxies can be described by only two numbers: an amplitude and a small deviation from scale invariance.

One of the predictions of the inflationary model is that there should be equal numbers of hot and cold spots in the microwave sky, and that the statistical properties of hot spots and cold spots should be identical. Analyses of both the WMAP and Planck satellite data reveal no evidence for this symmetry. Quantifying this through constraints on the three-point function, analyses show that the primordial fluctuations are symmetric to better than one part in a thousand. Any detection of these features would have been a significant challenge to the inflationary model. The statistical properties of these observations also show a remarkably strong agreement with the predictions of the inflationary scenario: the fluctuations are adiabatic, Gaussian, nearly scale-invariant, and coherent over scales that are larger than the "horizon" scale (the distance that light can travel). The statistical properties of the millions of points in the sky are described by two basic numbers, an amplitude and a scale-dependence – a remarkable success for the inflationary scenario.[14]

Another prediction of the inflationary model is that the geometry of the universe should be very close to flat. There should be nearly equal amounts of kinetic energy and gravitational energy. At the time that the inflationary universe was proposed, most astronomers would argue that the gravitational energy associated with the known galaxies was too small to balance the kinetic energy in expansion. In the 1980s, most cosmologists would have argued that the observations implied that the ratio of the two, usually called Ω, was 0.2 – 0.3. Inflationary models predicted Ω = 1. While a number of theorists noted the possibility that this discrepancy could be resolved if the universe was filled with dark energy, this possibility was considered exotic, the last refuge of the scoundrels who wanted to preserve the inflationary model. Today, the observational situation is very different. The WMAP and Planck satellite observations imply that Ω = 1 to better than 1 percent. When these cosmic microwave observations are combined with observations of large-scale structure, the current best measurements imply that Ω = 1 to better than 0.1 percent, another remarkable success for the inflationary model.

Despite these many predictive successes, the inflationary model faces a number of theoretical challenges. The inflationary scenario does not explain the origin of the universe and requires special initial conditions. For inflation to match the observed large size of the universe and the low amplitude of initial fluctuations, the parameters in the model must be fine-tuned. Since inflation occurs in the first moments of the Big Bang and at energy scales far beyond those accessible in the laboratory, our exploration of its physics stretches our basic understanding of the underlying nature of matter, space, and time.

Over the past few decades, cosmologists have developed a remarkably successful cosmology model that fits a host of astronomical observations. However, while this model addresses many of the previously unsolved questions of cosmology, it raises a new set of questions:

- Why is the universe accelerating? What is the nature of dark energy? Are we seeing the breakdown of gravity on cosmological scales?
- What is the nature of dark matter?
- Did the early universe also undergo a period of acceleration? If so, what was the mechanism that drove this early period of inflation?

There are many different routes toward addressing these questions. Developments in string theory and other attempts at unifying physics may provide new insights into the nature of space and time. Future observations of the geometry of the universe, the statistics of the primordial fluctuations, as well as the gravitational waves predicted in the inflationary scenario will either confirm this basic model or challenge its underlying tenets. Searches for dark matter could reveal the nature of these unknown particles. Astronomical measurements of distances or the growth rate of structure will test the notion that vacuum energy drives cosmic acceleration.

Of course, if we can address any of these questions, the answers will likely point toward even deeper and more profound mysteries.

ENDNOTES

[1] P. James E. Peebles, "Tests of Cosmological Models Constrained by Inflation," *The Astrophysical Journal* 284 (1984): 439–444; and Jeremiah P. Ostriker and Paul J. Steinhardt, "The Observational Case for a Low-Density Universe with a Non-Zero Cosmological Constant," *Nature* 377 (1995): 600–602.

[2] Adam G. Riess et al., "Observational Evidence from Supernovae from an Accelerating Universe and a Cosmological Constant," *The Astronomical Journal* 116 (1998): 1009–1038; and Saul Perlmutter et al., "Measurements of Omega and Lambda from 42 High-Redshift Supernovae," *The Astrophysical Journal* 517 (1999): 565–586.

[3] David N. Spergel et al., "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations; Determination of Cosmological Parameters," *The Astrophysical Journal Supplement* 148 (2003): 175–194.

[4] George F. Smoot et al., "Structure in the COBE Differential Microwave Radiometer First Year Maps," *The Astrophysical Journal* 396 (1992): L1–L5.

[5] Charles L. Bennett et al., "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Results and Basic Maps," *The Astrophysical Journal Supplement* 148 (2003): 1–27.

[6] Planck Collaboration; Peter A.R. Ade et al., "Planck 2013 Results, I. Overview of Products and Scientific Results," accepted by *Astronomy and Astrophysics*, doi:10.1051/0004-6361/201321529, arXiv:1303.5062.

[7] Erminia Calabrese et al., "Cosmological Parameters from Pre-Planck CMB Measurements," submitted to *Journal of Cosmology and Astroparticle Physics*, arXiv:1302.1841.

[8] Erik Aver, Keith A. Olive, and Evan D. Skillman, "An MCMC Determination of the Primordial Helium Abundance," *Journal of Cosmology and Astroparticle Physics* 4 (2012): 1–23; and Fabio Iocco et al., "Primordial Nucleosynthesis: From Precision Cosmology to Fundamental Physics," *Physics Reports* 47 (2009): 1–76.

[9] Adam G. Riess et al., "A 3% Solution: Determination of the Hubble Constant with the Hubble Space Telescope and Wide Field Camera 3," *The Astrophysical Journal* 730 (119) (2011): 1–18.

[10] A. Conley et al., "Supernova Constraints and Systematic Uncertainties from the First Three Years of the Supernova Legacy Survey," *The Astrophysical Journal Supplement* 192 (2011): 1–29.

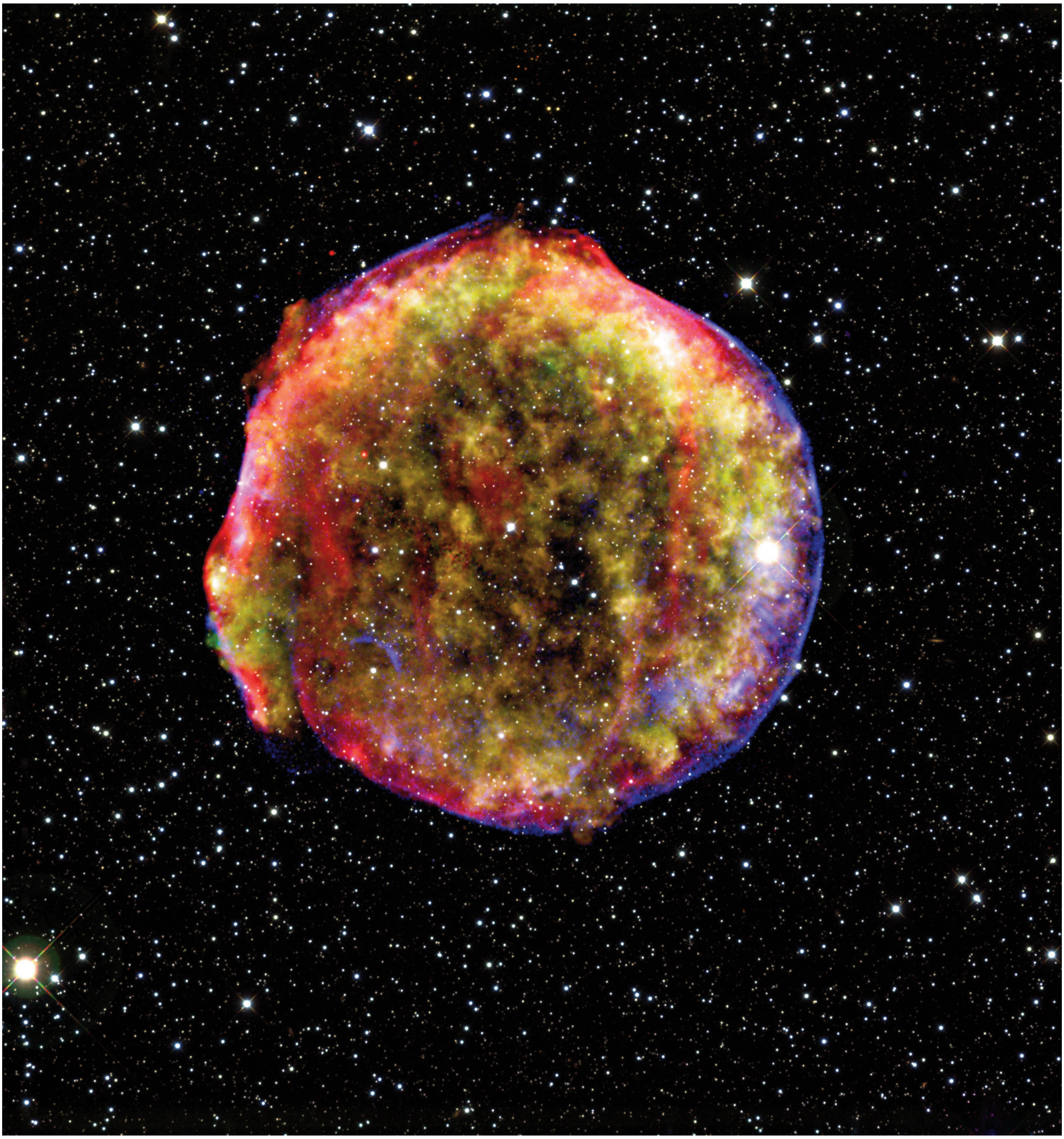
[11] Alexei Vikhlinin et al., "Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints," *The Astrophysical Journal* 692 (2009): 1060–1074.

[12] Beth A. Reid et al., "Cosmological Constraints from the Clustering of the Sloan Digital Sky Survey DR7 Luminous Red Galaxies," *Monthly Notices of the Royal Astronomical Society* 404 (2011): 60–85; and Lauren Anderson et al., "The Clustering of Galaxies in the SDSS-III Baryon Acoustic Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Releases 10 and 11 Galaxy Samples," *Monthly Notices of the Royal Astronomical Society* 441 (1) (2014): 24–62.

[13] Alan H. Guth and Paul J. Steinhardt, "The Inflationary Universe," *Scientific American* 250 (1984): 116–128; and Andrei Linde, "Particle Physics and Inflationary Cosmology," *Physics Today* 40 (1987): 61–68.

[14] Hiranya Peiris et al., "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Implications for Inflation," *The Astrophysical Journal Supplement* 148 (2003): 213–231; and Planck Collaboration; Peter A.R. Ade et al., "Planck 2013 Results, XXIV. Constraints on Primordial Non-Gaussianity," accepted by *Astronomy and Astrophysics*, doi:10.1051/0004-6361/201321554, http://planck.caltech.edu/pub/2013results/Planck_2013_results_24.pdf.

Inside back cover : Tycho's Supernova Remnant (SN 1572, G120.1+01.4). A composite image of X-ray (yellow, green, blue), infrared (red), and optical light (white stars) displaying the remnants of a Type Ia supernova in the constellation Cassiopeia more than four centuries after explosion. The supernova remnant is named after Tycho Brahe, the Danish Renaissance astronomer who recorded the event in 1572. The image is 15.5 arcmin across. X-ray : NASA/CXC/SAO ; Infrared : NASA/JPL-Caltech ; Optical : MPIA, Calar Alto, O.Krause et al. Composite image accessed at the Chandra X-Ray Observatory website, http://chandra.harvard.edu/Photo/2009/Tycho.



coming up in Dædalus:

What is the Brain Good For?	Fred H. Gage, Thomas D. Albright, Emilio Bizzi & Robert Ajemian, Brendon O. Watson & György Buzsáki, A. J. Hudspeth, Joseph LeDoux, Earl K. Miller & Timothy J. Buschman, Terrence J. Sejnowski, Larry R. Squire & John T. Wixted, and Robert H. Wurtz
On Water	Christopher Field & Anna Michalak, Michael Witzel, Charles Vörösmarty, Michel Meybeck & Christopher L. Pastore, Terry L. Anderson, John Briscoe, Richard G. Luthy & David L. Sedlak, Stephen R. Carpenter & Adena R. Rissman, Jerald Schnoor, and Katherine Jacobs
On an Aging Society	John W. Rowe, Jay Olshansky, Julie Zissimopolous, Dana Goldman, Robert Hummer, Mark Hayworth, Lisa Berkman, Axel Boersch-Supan, Dawn Carr, Linda Fried, Frank Furstenberg, Caroline Hartnett, Martin Kohli, Toni Antonucci, David Bloom, and David Canning
Food, Health & the Environment	G. David Tilman, Walter C. Willett, Meir J. Stampfer & Jaquelyn L. Jahn, Nathaniel D. Mueller & Seth Binder, Steven Gaines & Christopher Costello, Andrew Balmford, Rhys Green & Ben Phalan, G. Philip Robertson, Brian G. Henning, and Steven Polasky

plus The Internet; What's New About the Old?; New Dilemmas in Ethics, Technology & War &c

American Academy
of Arts & Sciences
Cherishing Knowledge · Shaping the Future

U.S. $13; www.amacad.org