DATA-RICH ORGANIC CHEMISTRY

Enabling and Innovating the Study of Chemical Reactions

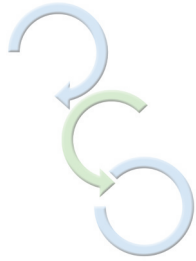
AN NSF-SPONSORED WORKSHOP

DATA-RICH ORGANIC CHEMISTRY

Enabling and Innovating the Study of Chemical Reactions

AN NSF-SPONSORED WORKSHOP
September 11-12, 2014, Washington, DC

Prepared by
Donna G. Blackmond, Ph.D. and Nick Thomson, Ph.D.



This material is based on work supported by the National Science Foundation under grant CHE-1447743. Any opinions, findings and conclusions or recommendations expressed in this material are those of the participants and do not necessarily reflect the views of the National Science Foundation.

Contents

Executive Summary

The purpose of this workshop was to bring together a group of academic, industrial, and governmental scientists to explore the challenges and opportunities inherent in the modern study of complex organic reactions related to the pharmaceutical industry. As our ability to monitor reactions in real time expands, and as the accuracy and data-density of our measuring tools increase, we face the challenge of developing an integrated approach to data capture and interpretation. The need to build new collaborative funding models to exploit innovations in the study of chemical reactions follows from these challenges. The need for a review of how our educational programs should exploit these significant developments was highlighted as we aim to develop an efficient and productive workforce for the 21st century.

The workshop was attended by a diverse group of 29 academic chemists and chemical engineers, 19 scientists and engineers in the pharma, analytical instrument, and related industries, and 5 government/nonprofit employees (App. 1). The workshop agenda began with a discussion of the current state of collaborative research between academia and industry, followed by a session outlining the opportunities for transformative pharmaceutical solutions leading to precompetitive collaborations. Key challenges were highlighted, including the need to develop a coherent vision for a common data framework, to understand the landscape for new data-rich technologies, and to align future priorities from industry, academia, and government perspectives, as well as to brainstorm new blue-sky horizons. One session focused on how we can incorporate modern data-rich tools into our undergraduate and graduate educational programs, both for accelerating breakthrough academic research and for preparing a productive future workforce. Workshop participants engaged in a succession of discussion stations focusing on topics including dissemination of data-rich tools and methodologies; developing the optimal collaborative models; defining the challenges that may be met with data-rich tools; and developing novel educational approaches and new tools of broad use to the general community.

The final focus of the workshop was in defining the path forward. A general consensus settled on two issues:

I. A proposal for a **new educational model**. Led by Clark Landis, with the core foundation of industrial case studies (App. 2)

II. A proposal for a **new networked core facility for data-rich experimentation** (App. 3), learning and extending from the Caltech model, enabling more rapid and in-depth development of academic chemistry as well as providing training in data-rich methodologies. Each of these ideas will form the basis of a sharply focused NSF workshop proposal to be submitted within the next year.

Introduction

In June 2013, The Council for Chemical Research (CCR) hosted a New Industrial Chemistry & Engineering (NIChE) workshop on the topic of "Precompetitive Collaborations on Enabling Technologies for the Pharmaceutical Industry" at the University of Pennsylvania. Chaired by three pharmaceutical scientists: Christopher Welch (Merck), Joel Hawkins (Pfizer) and Jean Tom (Bristol-Myers Squibb), the meeting brought together leaders in industry, academia and government to explore new approaches to cross-pharma collaborations on precompetitive chemistry and chemical engineering technologies.

This workshop highlighted significant recent advances in data-rich measurement capabilities in the pharmaceutical industry. The growing need for rapid information collection in an era of shrinking resources provides a strong motivation for pre-competitive collaboration between companies themselves and between companies and academia. One main theme arising from the CCR workshop was the need for a strong focus on sustainability as we seek an integrated approach to data capture and interpretation. How can we best implement new technologies in a "Lab of the Future" to streamline process research and development through fundamental process understanding? How can we intelligently navigate big data for clarity rather than confusion? The questions raised in these discussions led to the proposal for the current workshop. NSF sponsorship is critical to an expanded academic perspective to explore these ideas in further detail.

The broad aim of this workshop is to drive sustainability of the US economy and workforce through dissemination of data-rich tools across industry and academia, through the building of new collaborative funding models across academia, industry and government, and through the implementation of ideas for the further development of our workforce.

Two principal outcomes emerged from this NSF workshop: i) a proposal for the development of a data-rich Experimentation Center, spearheaded by Joel Hawkins (Pfizer); and ii) a proposal for a new educational model with industrial case studies highlighting data-rich issues championed by Clark Landis (Wisconsin). Both ideas aim to be articulated as formal proposals for NSF workshops to be submitted within the next calendar year.

Session I

OPENING: BACKGROUND AND SCOPE OF THE WORKSHOP

Nick Thomson (Pfizer) opened the workshop with an introduction that set the stage for the two days of discussions. Following on from the 2013 UPenn workshop on pre-competitive collaboration sponsored by the CCR, the strong focus on sustainability by collaboration across industry, academia, and government was reiterated. The "Lab of the Future" is generally seen as the way to secure this sustainability by learning to navigate voluminous data sets for clarity rather than confusion. The dissemination of data-rich tools across academia and industry, the building of new collaborative models for future innovations, and to develop our future workforce in a data-rich environment, were all highlighted as aims for this workshop.

The outputs from this workshop will include this comprehensive report as well as concrete proposals for future projects to exploit further the ideas developed in these discussions.

Session II

THE CURRENT STATE

Understanding the construct. This session outlined a variety of models currently in place for academic, industry and government engagement in data-rich experimentation. The presentation of each case study and the discussion sessions that followed were moderated by Donna Blackmond.

A. CCHF – Center for Selective C-H Functionalization (Huw Davies, Emory)

The CCHF is an NSF CCI (Center for Chemical Innovation) currently in Phase 2, involving 13 academic and several industrial partners. Its mission statement is defined by chemistry development that is "revolutionary rather than evolutionary" with a goal of industrial engagement to realize the commercial potential of the powerful synthetic chemistry and catalytic methods developed via this Center. The Center has a strong interdisciplinary, multi-laboratory, international focus. The use of state-of-the-art video conferencing tools is an important hallmark of the success of this Center, including weekly meetings of the various themed subgroups in the Center, laboratory exchanges, and distance learning programs. Outreach, diversity, and broader impact of the Center's work have been documented. Industrial

involvement is offered on several different levels, ranging from pre-publication access to discoveries to high level collaborating partnerships. Case studies from each level of industrial involvement were discussed. Discussion ensued of issues such as how the Center deals with IP issues. The most successful involvement has been in the pre-competitive space.

B. 3CS – Caltech Center for Catalysis and Chemical Synthesis (Scott Virgil, Sarah Reisman, Caltech)

This Center, initiated with private funding, brings together state-of-the-art robotics and high-throughput analytical instrumentation with a substantial archive of catalysts/ligands and highly trained personnel to promote catalyst discovery and mechanistic advances, with an educational component enabled by a walk-up facility and library design protocol. Data storage, management, and software workflow as well as hardware maintenance are key issues in keeping the center sustainable. This core facility showcases the power of high throughput instrumentation in analyzing a variety of reaction processes. This case study was viewed by workshop participants as a powerful model that could be emulated in future data-rich applications.

C. Merck NSF-GOALI Experience (Shane Krska, Merck)

The Merck Catalysis Laboratory is a centralized facility focusing on the application of cutting-edge catalytic methods to address process and medicinal chemistry problems. Significant infrastructural support for high-thoughput experimentation as well as a mandate to interface with academic experts exists. Three case studies were discussed in detail: i) Gary Molander, U Penn; ii) Mitch Smith, Michigan State; iii) Paul Chirik, Princeton. In the first case, what started as a no-cost, informal collaboration, with Penn postdocs spending time at Merck labs to learn high throughput techniques, ultimately culminated in the building of a GOALI-funded HTE Center at Penn that is now self-sustaining and has resulted in over 25 joint publications. The organic development of the collaborations in each case, with preliminary results produced informally, led naturally to the GOALI proposals, which have turned out to be a powerful model to fund these partnerships and promote academic-industrial synergy that has been an invaluable education tool for the students and post docs involved. Academic outgrowths of the collaborations have included applications to undergraduate teaching. The GOALI model is also noted as a powerful pump-priming method for future NSF proposals from the academic partners. Significant value exists in bringing together academic and industrial perspectives to enable key scientific discoveries in areas critical to industry. Subsequent discussion later in the workshop favored the option of multi-pharm with academic GOALI grants.

D. SSPC – Solid State Pharmaceutical Cluster (Joe Hannon, Dynochem)

The SSPC represents a highly focused governmental investment in R&D on the solid state properties of active pharmaceutical intermediates (API) by Ireland. The Center explores techniques in data-rich measurement of pre-competitive problems in three strands: synthesis, crystal growth, and drug product formulation. Key outputs

include publications, engagement aimed at job sustainability, spinoff CROs, and sharing of technology. The need to insure a good balance between basic and applied research was noted. Online tools are extensively used to help the collaborative efforts. The challenges inherent to an industrial case study help to ensure pragmatic focus of the partners and very strong government financial support ensures a program of significant scale. Other notable efforts in the EU include CMAC in the UK and RCPE in Austria.

E. UK Pharmacat Model (Mimi Hii, Imperial College)

The Pharmacat consortium was conceived as a flexible grant scheme funded by pharma company members to promote interdisciplinary research between academic chemists and chemical engineers at Imperial College. The pharma scientists outline in broad terms the general areas of interest, both short- and long-term (some have been delineated by the ACS Green Chemistry Initiative). Academic scientists submit proposals for funding, which are reviewed by a core group of the industrial partners. Each funded project is assigned an industrial lead, and meetings ca. every three months evaluate progress. Student visits to industrial labs to carry out research are encouraged. Six-monthly update meetings attended by all participants of the Scheme highlight the research accomplishments over the year. The grants may serve as seed money for future proposals to government funding agencies, thus leveraging the original pharma investment more than three-fold. Strong research collaborations between academia and industry are promoted by this model.

Recent progress in pre-competitive collaboration. Reasons for the considerable recent interest in pre-competitive collaborations on enabling technologies across the pharma industry was examined. It was noted that the development of new technologies can be highly inefficient if done alone; the pooling of resources on problems of common interest and the leveraging of shared ideas, inputs, and feedback can have a multiplicative effect among collaborators. The three presenters have co-authored a paper (OPRD, 2014, 18, 481) that arose from the discussions at the 2013 CCR meeting. Here they discussed current progress in the Lab of the Future, process chemistry, engineering and analytical science.

A. (Joel Hawkins, Pfizer)

While an overarching goal is to enable maximum cross-pollination of ideas, potentially including multiple industrial, academic, and vendor partners, it was recognized that not all enabling technologies fit this pre-competitive model, and antitrust safeguards acceptable to all parties need to be in place. The former model of pharma development, with its lack of basic process understanding and slow iterative scale-up, is widely seen to be no longer sustainable, given pressures to accelerate development. The quality landscape has also changed significantly, and the quality by design approach that is gaining acceptance requires deep process understanding. The Lab of the Future at Pfizer brings a suite of data-rich tools, which may be managed by the analyst, the chemist, and the chemical engineer. The need to disseminate these tools is becoming even more urgent. Data needs to be transportable – across people, across time, across location. Broad utilization requires appropriate software, capable of facile data integration and visualization. Tools using a common language can facilitate exchange between generalist and specialist cultures. Much concern over consolidation of pharma and job losses in recent years motivates the drive towards innovation and sustainability.

B. (Jean Tom, BMS)

From the engineering perspective, data-driven process development requires that all steps in the overall process are considered together. Screening, high throughput DOE, data visualization and analysis, and model development all contribute to data-driven process development. The key problem is the integration of data into searchable architecture.

C. (Chris Welch, Merck)

New enabling technologies that have been appearing in recent years need to be evaluated; The "First at Merck" efforts to identify, acquire an evaluate all new tools (hardware and software) were described that have the potential to accelerate process development. These have included general lab equipment, reagents and catalysts, purification/separation technologies, and data-rich experimentation/monitoring tools. In some cases this has led to the joint development of a tool between Merck and a vendor. Examples of innovative analytical equipment developed in this way were discussed, including a small mass spec detector. Ultimately, the goal is to provide data-rich tools without data handling headaches.

The potential next steps in this approach were discussed, including a multi-pharma GOALI idea and a DARPA-type model for government funding to encourage innovation.

Session III

THE OPPORTUNITY: TRANSFORMATIVE PHARMA SOLUTIONS (Nick Thomson, Pfizer)

It is widely agreed that the traditional way of optimizing reactions – round bottom flasks set up for overnight runs with a TLC analysis the next morning – have always been problematic for efficient scale-up and are not sustainable in the current environment. The acceleration of drug development comes from many quarters, including the FDA's "breakthrough therapy" designation. The new paradigm of precision medicine has evolved to the point where drug candidates can be taken directly from Phase 2 to the market. Process development must be prepared for such acceleration.

The quality landscape of current and future FDA culture requires a harmonized pharmaceutical platform applicable across the lifecycle of the product emphasizing an integrated approach to quality risk management and science. Data-rich measures of quality can help to accelerate development and build in quality from the outset of the process development. This changing environment makes a strong case for

the concept of the Lab of the Future. Evolving laboratory technologies are driving many positive developments in data-rich reaction monitoring. Carefully controlled reactors with myriad probes provide detailed process understanding – an important outcome both in industry, where this information is critical to scale-up, and in academia, where mechanistic understanding leads to new catalyst discovery and reaction optimization. Integration of all of the available data remains a challenge, as does the need to disseminate these tools – and how to use them – across academia. Innovation in experimental tools along with new synthetic methodologies, new ancillary technologies, computational prediction methods, and data management architecture all contribute to the aim to ensure sustainability.

New skills will be required to prepare our worforce for this data-rich world of the Lab of the Future; intelligent collaboration will be an additional drive to intelligent deployment of big data techniques. Will new funding mechanisms be needed to foster these collaborations?

Panel Discussion With Session II and III Speakers.A general discussion between all workshop participants and the speakers of Sessions II and III followed. Several academic participants noted that many universities encounter problems with IP issues in trying to develop industrial collaborations. Huw Davies commented on the CCHF experiences with a multi-university, multi-company, NSF-sponsored consortium, which could easily become unwieldy but is made simpler by having a basic, standard agreement in place, as well as taking care in virtual meetings to note when the meeting is “open” and when it is “closed.”

A discussion of the different models for collaboration presented in Session II prompted the suggestion of an analogy to the successful consortia in the semiconductor industry in the 1980’s. Originally funded by the DoD, these became self-sustaining over time. Some of the same driving forces apply in pharma. The Fraunhofer Institute was mentioned as an example for industrial innovation that must eventually become self-sustaining.

A discussion ensued about various models for how basic research inspires applied developments. Is basic science the driver for technology? Is use-inspired research a better (or at least equivalent) driver? Does technology also drive directions in basic science? Is truly disruptive technology ready for “prime time” at its outset? Should NSF include more industrial scientists in its reviewing processes, since they may be better placed to make the case for potential applicability of emerging technologies? It was also noted, however, that NSF Chemistry Division may not be the most appropriate place for applied science proposals.

The question of economics was raised: what are the areas of research where the greatest impact can be made? What is the economic impact of data-rich process technology? For academics coming from outside pharmaceutical connections, how to find out about the best “real” problems that require basic research input? The Green Chemistry Roundtable was noted as one venue where pharma researchers were able to articulate the biggest fundamental science challenges they face.

Data-rich experimentation fits into both academic and industrial laboratories. Everyone wins when this is developed collaboratively. It would be helpful if industry could take a longer-term view. Forming broad partnerships often leads to broader use of ideas.

Session IV

KEY CHALLENGES

A. Developing a Common Data Framework (D. Vanderall, BMS)

The increasing use of data-rich experimentation means that we are increasingly being buried under data! The root problem lies in the lack of data standards. The sharing and mining of data is hindered by this lack of connectivity. The Allotrope Framework was developed to address this issue. Allotrope was founded by a consortium of pharma companies including Merck, Pfizer, Amgen, and Abbvie.

There are a number of "choke points" in data analysis and sharing. These include document preparation, extracting knowledge and value from data, limited data exchange, data management and archiving, dealing with errors, and regulatory compliance. The causes of these problems stem from a lack of standard file formats, incomplete/incompatible software, and inconsistent metadata. Creating a common data framework as a toolkit that is independent of the technique or the vendor is under development.

B. The Landscape for Developing New Technologies (H. Dubina, Mettler Autochem)

New technologies may be developed in joint projects that are living and evolving organisms, with these stages: planning, pilot, implementation, sustainability, refinement, further developments. Government leadership has to date been stronger in the UK and Singapore compared to the U.S.

We need to identify the gaps that exist between ideas and execution that can be filled through collaborations between tech partners, industry, and academia. The challenges of joint development include ensuring that the partners develop a coherent vision that manages all the stakeholders' priorities. Smaller scale, more agile projects appear to be best suited for success.

C. Future priorities from an industry perspective (M. Faul, Amgen)

The IQ consortium collaborations is composed of 37 companies, with the purpose to advance science-based and science-driven standards and regulations. Because many of the members are competitors, the consortium must maintain strict compliance with anti-trust laws.

"Pre-competitive" in this context is defined as a collaboration between two or more pharma companies that is designed to produce an efficiency enhancing advancement in which no company has a competitive interest. The Enabling Technologies Work Group within IQ was set up to identify gaps in enabling technologies within pharma and foster pre-competitive collaborations between the pharma members to address these gaps. Optimization of technologies that enable pharma development to increase efficiency, while maintaining IP protection, is a key challenge. This challenge may be further broken down into: i) establishing what is pre-competitive; ii) defining the models that best reflect the opportunities; iii) developing and delivering technology via openly established and transparent collaborations; iv) gaining small wins for big successes (starting with smaller "proof of principle" opportunities to demonstrate success); v) delivering results quickly via efficient collaborations (agreeing on a timeline and assigning accountable points of contact; understanding all partners' goals); vi) defining IP requriements and expectations in academic collaborations.

Breakout Discussion: How can we develop a sustainable platform of technologies to better integrate our workflows and convert data into knowledge more seamlessly? See App. 8, Participant Notes, for more details.

Session V

BLUE SKY CHALLENGES

What questions might we be able to address by the parallel advances in data-rich analytical, theoretical, and computational sciences that we are not able to address at present? A brainstorming session on new horizons for data-rich chemistry attempted to answer this question.

Scott Miller spoke about "Grand Challenges and Holy Grails", discussing organic chemistry beyond the functional group ("post Morrison and Boyd"). Some topics include late-stage C-H functionalization, combinatorial catalysis, development of assays for complex mixtures, modification of complex molecules. Matt Sigman discussion prediction in science (analogous to the same in society) using big data. Key points are the parameterization of organic chemistry, the use of experimental design, and the development of complex models that relate back to mechanism, and may be used for prediction of phenomena such as site selectivity. Clark Landis spoke about the development of NMR reactor and the development of robust kinetic models that may be used to demonstrate our understanding of the reaction: "ab initio full kinetic modeling" as a goal. The challenge of big data in process systems engineering and design, specifically in real-time decision and control for smart manufacturing, was highlighted by Wayne Bequette. Philip Hopke discussed multivariate curve resolution and evolving factor analysis; organic chemists need to think harder about models for our reaction data and include ways to look at time-variant systems.

Session VI

EDUCATING TOMORROW'S WORKFORCE

A panel discussion provided the format for addressing the question of bringing the Lab of the Future to the classroom. A key point is to define the required workplace skill sets for future generations. What should we provide at the undergraduate level? How can we improve the student experience with industrial ideas and problems? Time spent in industry (i.e. internships) is generally agreed to be extremely valuable, but insufficient funds exist at present to support these for a significant number of students. The main points of the panel members are summarized here.

Opportunities

Data-rich approaches to chemistry research are basically missing from our undergraduate and graduate curriculum. There are significant opportunities to develop new teaching laboratories and new coursework that will develop critical skills in data-rich science. One leveraging opportunity is to develop curricula, problems and laboratories based on industry case studies and data sets, to make meaningful connections with industrial research.

There is a great desire among students and faculty for more internships and professional development opportunities for students considering industrial careers.

Session VII

LEARNING AND DISCUSSION STATIONS

How do we disseminate current tools? (H. Dubina, Mettler Autochem)

Henry Dubina from Mettler led a session on how to best disseminate current tools. The cost of current instrumentation can be somewhat prohibitive. A core idea was the opportunity to better leverage high throughput data rich experimentation centers, similar to the Cal Tech model. The ability to develop a shared pool of instruments might allow us to lower cost and improve access, while recognizing the different requirements of industry and academia. Developing mechanisms to repurpose older equipment for research purposes, bringing it up to the state of the art, would also be beneficial to the community.

What is the best collaborative model going forward that will meet our different needs? (M. Faul, Amgen)

Margaret Faul from Amgen led a session on the best collaborative models to meet our different needs. There was significant focus on developing our networks as a spring board to future collaboration, using social media, conferences, round tables and other networking events. A number of factors were discussed as being critical for success, including well-defined deliverables, clarity on funding options and availability, strong communication, multi-discipline engagement, integrated student learning opportunity, flexibility, passionate members, and access to real life examples and substrate.

What are the big challenges that we can resolve with data driven tools? (K. Jensen, MIT)

Klavs Jensen from MIT led a session on what big challenges can be resolved with data driven tools. There was strong agreement on the desire to use such tools to drive depth of understanding, knowledge and learning, to capture historical data, to mine both successes and failures, in order to drive new insight on reactivity and structure, and greater parameterization of physical organic chemistry. This in turn should allow us to discover new chemistries and technologies. A key element to be resolved is the need for improvement in data (and meta data) capture combined with mechanisms to make it readily available to the community, for example through a broad and common electronic laboratory notebook platform. As we improve data access, there is significant opportunity to develop artificial intelligence, machine learning and more autonomous development.

What new tools might be of broad use to the community? (S. Tummala, BMS)

Srinivas Tummala from BMS led a session on new tools that would be of interest to the community. The most needed tools were categorized in to data collection/ capture, analysis and archiving. First of all, we should look to derive as much benefit from the current suite of available tools, through models that lower cost, improve access, prolong lifespan and broadly share success stories across the community. In the next generation of tools, balancing sophistication with ease of use will be key to adoption, as will the ability to lower cost through more robust hardware and common, open source software platforms and data archiving systems.

How can we develop novel educational approaches? (J. Hein, UC Merced)

Jason Hein and Clark Landis led a session on the development of novel educational approaches to train students for future careers. It was agreed that students should be exposed to real problems that reflect modern research. New collaborative educational models are required to train the art of multi-disciplinary problem solving, provide access to instrumentation and provide data rich real life examples.

Session VIII

DEFINING THE PATH FORWARD

Further discussion on the learning station outcomes led the group to decide to focus on four key ideas:

- Development of new educational models (Clark Landis)
- Development of a 'Cal Tech like' data rich experimentation hub (Joel Hawkins)
- Development of new industrial/academic collaboration models (Shane Krska)
- Development of future grand challenges to be addressed through data rich experimentation (Klavs Jensen)

Of these four ideas, the first two listed – the development of a new educational model and the development of a data rich experimentation hub – were prioritized as the most important to take to the next stage.

Appendices

Appendix 1

List of Workshop Participants

Name	Affiliation
Ryan Baxter	University of California, Merced
Wayne Bequette	Rensselaer Polytechnic Institute
David Berkowitz	University of Nebraska–Lincoln
Steve Bernasek	National Science Foundation
Josh Bishop	Perkin Elmer
Donna Blackmond	The Scripps Research Institute
Nicholas Brunelli	Ohio State University
Silas Cook	Indiana University
Kathy Covert	National Science Foundation
Huw Davies	Emory University
Abigail Doyle	Princeton University
Henry Dubina	Mettler Toledo AutoChem
Martin Eastgate	Bristol-Myers Squibb
Margaret Faul	Amgen
Anne Fischer	DARPA
Fraser Fleming	National Science Foundation
David Ford	Nalas Engineering
Ramesh Giri	University of New Mexico
Carlos Guerrero	University of California, San Diego
Joe Hannon	DynoChem
Ryan Hartman	University of Alabama
David Harwell	American Chemical Society
Joel Hawkins	Pfizer
Jason Hein	University of California, Merced
Mimi Hii	Imperial College London

Philip Hopke	Clarkson University
Klavs Jensen	Massachusetts Institute of Technology
Michael Jones	Waters
John Kitchin	Carnegie Mellon University
Rob Knowles	Princeton University
Shane Krska	Merck
Clark Landis	University of Wisconsin Madison
Scott Miller	Yale University
Sharon Neal	University of Delaware
Timothy Newhouse	Yale University
Alex O'Brien	GlaxoSmithKline
Chuck Orella	Merck
Sarah Reisman	California Institute of Technology
Rob Rioux	Pennsylvania State University
Jerry Salan	Nalas Engineering
Charles Santa-Maria	Pfizer
Susannah Scott	University of California, Santa Barbara
Chris Senanayake	Boehringer-Ingelheim
Matt Sigman	University of Utah
Ryan Stowe	Scripps
Jose Tabora	Bristol-Myers Squibb
Nick Thomson	Pfizer
Jean Tom	Bristol-Myers Squibb
Srinivas Tummala	Bristol-Myers Squibb
Dana Vanderwall	Bristol-Myers Squibb
Scott Virgil	California Institute of Technology
Chris Welch	Merck
Sheryl Wiskur	University of South Carolina
Steve Wittenberger	Abbvie

Appendix 2

Proposal for New Educational Model

Clark Landis
University of Wisconsin

In the course of our workshop, we became aware that some pharmaceutical companies already use case studies and data sets for specific training needs. We propose that existing industry-based case studies constitute ideal thematic cores from which new pedagogical "data rich" chemistry materials can be developed. Because the case studies come from real-world chemistry discovery and development projects they comprise activities spanning high throughput reaction screening, real-time reaction monitoring, design of experiments, kinetic analysis of reaction mechanism, process economics and engineering, and archiving, retrieving, and analyzing large sets of data. Our goal is to create classroom, laboratory, and research activities at levels of sophistication that range from second-year undergraduates to postgraduate studies and involve all the traditional subdisciplines (organic, inorganic, analytical, physical) of chemistry and likely crossing into allied fields of study (medicinal chemistry, informatics, chemical engineering, etc.) while retaining a focus on organic chemical transformations. Development efforts will involve professionals from academia (primarily undergraduate institutions through research universities), industry (large pharma, chemical, and instrumentation companies through small businesses), and societies such as the American Chemical Society.

The number one conclusion of the ACS Presidential Commission report, Advancing Graduate Education in the Chemical Sciences, states "Current educational opportunities for graduate students, viewed on balance as a system, do not provide sufficient preparation for their careers after graduate school. " Preparation of students for future careers has many elements, including awareness of the real problems, techniques, skills, and people deployed in modern chemistry research. Our emphasis on authentic chemistry and data rich methods tackles career training needs in an environment of rapidly expanding information sources.

To promote the training of today's students for tomorrow's careers, we plan to submit a proposal to NSF that supports development of pedagogical materials centered on a few (three to five) case studies.

Appendix 3

Proposal for a Small Focused Scoping NSF Workshop on Extending the Caltech Model for Data Rich Experimentation in Academic Chemistry

Joel M. Hawkins
Pfizer Worldwide R&D

At the NSF Workshop on Data Rich Experimentation in September 2014 we discussed the Caltech model and potential ways to extend or expand this model. Dr. Scott Virgil and Prof. Sarah Reisman described the Center for Catalysis and Chemical Synthesis at Caltech which was set up initially with private money to provide high throughput reaction screening and quality analytics to the Chemistry Department, see: http://www.cce.caltech.edu/content/center-catalysis-and-chemical-synthesis-3cs. This Center is much more than a service group. Dr. Virgil works very closely with the students and postdocs to apply high throughput reaction screening to their chemistry projects such as catalyst development, thus linking the hardware and software of automation with the chemistry of the individual PIs, and Dr. Virgil adopts the automation as needed to serve new problems in new ways. As a result:

- The chemistry of individual PIs is discovered and developed faster and/or in greater depth.
- Students and postdocs learn to experiment and think in a high throughput and data rich fashion.

We had a breakout session exploring ways to extend or expand the Caltech model and proposed the following approach:

- Set up a second center at a university, potentially on the east coast in proximity to other universities and pharmaceutical companies. Maintain the features described above where the chemistry of individual PIs is discovered and developed faster and/or in greater depth and students and postdocs learn to experiment and think in a high throughput and data rich fashion.
- Promote communication of the new center with the Center at Caltech and existing and nascent academic automation groups (e.g. at the University of Pennsylvania, the University of Illinois, and UC Berkeley) in order to share best practices and new approaches.
- Further promote communication and collaboration of these academic groups with high throughput screening groups in the pharmaceutical industry to share and promote best practices with industry and to promote precompetitive collaboration with and within the industry.
- Incorporate in situ analytics for reaction profiling and kinetic modeling, thus

bringing in the complementary approaches of the in depth study of individual reactions with the screening of many reactions.

• Promote communication and collaboration with reaction engineering groups in the pharmaceutical industry, in particular for reaction profiling and kinetic modeling, in order to share best practices and new approaches to data rich experimentation.

• Expose the students and faculty to industrially relevant problems through this communication with industry.

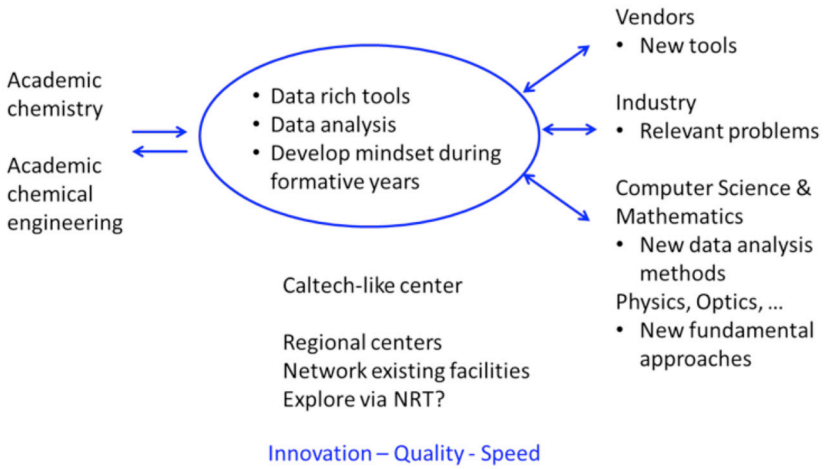
• Promote interactions of the center with academic groups developing computational methods, including those directed at other applications which could be applied to the statistical and kinetic analysis of high throughput chemical experimentation.

• Promote interactions of the center with academic groups in fields such as analytical chemistry, physics, and optics to convey needs and opportunities for reaction analysis, and to explore the application of new measurement techniques, including those directed at other applications.

• Promote interactions of instrument vendors with the new center, e.g. to test prototype instruments or to ultimately commercialize new measurement techniques studied in the center.

• Encourage secondments into the center from other universities and industry to share and disseminate data rich technologies.

• Work within the center on chemical problems from other universities.



Here we propose a small NSF workshop to scope out this idea further. The broadest goals of the proposed center include promoting sustainability and competitiveness through cross-disciplinary education and through the enhanced speed and quality of research and development in our industries.

Appendix 4

Links to Funding Resources

Research Opportunities:

GOALI — Grant Opportunity for Academic Liaison with Industry.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504699)

This NSF-wide program promotes university-industry partnerships by making project funds or fellowships/traineeships available to support an eclectic mix of industry-university linkages. Special interest is focused on affording the opportunity for:

- Faculty, postdoctoral fellows, and students to conduct research and gain experience in an industrial setting;
- Industrial scientists and engineers to bring industry's perspective and integrative skills to academe;
- Interdisciplinary university-industry teams to conduct research projects.

This solicitation targets high-risk/high-gain research with a focus on fundamental research, new approaches to solving generic problems, development of innovative collaborative industry-university educational programs, and direct transfer of new knowledge between academe and industry. GOALI seeks to fund transformative research that lies beyond that which industry would normally fund.

CDS&E — Computational and Data-Enabled Science and Engineering.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504813&org=CHE&from=home)

This program works with other programs to enable CDS&E. Advanced computational infrastructure and the ability to perform large-scale simulations and accumulate massive amounts of data have revolutionized scientific and engineering disciplines. The goal of the CDS&E program is to identify and capitalize on opportunities for major scientific and engineering breakthroughs through new computational and data analysis approaches. The intellectual drivers may be in an individual discipline or they may cut across more than one discipline in various Directorates. The key identifying factor is that the outcome relies on the development, adaptation, and utilization of one or more of the capabilities offered by advancement of both research and infrastructure in computation and data, either through cross-cutting or disciplinary programs.

The CDS&E program welcomes proposals in any area of research supported through the participating divisions that:

- Promote the creation, development, and application of the next generation of mathematical, computational and statistical theories and tools that are essential for addressing the challenges presented to the scientific and engineering communities by the ever-expanding role of computational modeling and simulation and the explosion and production of digital experimental and observational data.

- Promote and encourage integrated research projects that create, develop and apply novel computational, mathematical and statistical methods, algorithms, software, data curation, analysis, visualization and mining tools to address major, heretofore intractable questions in core science and engineering disciplines, including large-scale simulations and analysis of large and heterogeneous collections of data.

- Encourage adventurous ideas that generate new paradigms and that create and apply novel techniques, generating and utilizing digital data in innovative ways to complement or dramatically enhance traditional computational, experimental, observational, and theoretical tools for scientific discovery and application.

- Encourage ideas at the interface between scientific frameworks, computing capability, measurements and physical systems that enable advances well beyond the expected natural progression of individual activities, including development of science-driven algorithms to address pivotal problems in science and engineering and efficient methods to access, mine, and utilize large data sets.

Supplement requests to existing awards within a program that address one of the points above will also be considered.

Chemistry: CDS&E encourages innovative and adventurous ideas that generate new paradigms at the algorithmic, software design and data acquisition levels in computational chemistry, simulations, chemical data analysis and cheminformatics, producing new ways of "doing business".

IUCRC – Industry/University Collaborative Research Centers.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5501)

The Industry/University Cooperative Research Centers (I/UCRC) program develops long-term partnerships among industry, academe, and government. The centers are catalyzed by a small investment from the National Science Foundation (NSF) and are primarily supported by industry center members, with NSF taking a supporting role in the development and evolution of the center. Each center is established to conduct research that is of interest to both the industry members and the center faculty. An I/UCRC contributes to the nation's research infrastructure base and enhances the intellectual capacity of the engineering and science workforce through the integration of research and education. As appropriate, an I/UCRC uses international collaborations to advance these goals within the global context.

CCI – Centers for Chemical Innovation.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13635&org=CHE&from=home)

The Centers for Chemical Innovation (CCI) Program supports research centers focused on major, long-term fundamental chemical research challenges. CCIs that address these challenges will produce transformative research, lead to innovation, and attract broad scientific and public interest. CCIs are agile structures that can respond rapidly to emerging opportunities and make full use of cyberinfrastructure to enhance collaborations. CCIs may partner with researchers from industry, government laboratories and international organizations. CCIs integrate research, innovation, education, and informal science communication and include a plan to broaden participation of underrepresented groups.

STC - Science and Technology Centers
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5541&org=)

The Science and Technology Centers (STC): Integrative Partnerships program supports innovative, potentially transformative, complex research and education projects that require large-scale, long-term awards. STCs conduct world-class research through partnerships among academic institutions, national laboratories, industrial organizations, and/or other public/private entities, and via international collaborations, as appropriate. They provide a means to undertake significant investigations at the interfaces of disciplines and/or fresh approaches within disciplines. STCs may involve any area of science and engineering that NSF supports. STC investments support the NSF vision of creating and exploiting new concepts in science and engineering and providing global leadership in research and education.

Centers provide a rich environment for encouraging future scientists, engineers, and educators to take risks in pursuing discoveries and new knowledge. STCs foster excellence in education by integrating education and research, and by creating bonds between learning and inquiry so that discovery and creativity fully support the learning process.

NSF expects STCs to demonstrate leadership in the involvement of groups traditionally underrepresented in science and engineering at all levels (faculty, students, and postdoctoral researchers) within the Center. Centers use either proven or innovative mechanisms to address issues such as recruitment, retention and mentorship of participants from underrepresented groups.

Centers must undertake activities that facilitate knowledge transfer, i.e., the exchange of scientific and technical information with the objective of disseminating and utilizing knowledge broadly in multiple sectors. Examples of knowledge transfer include technology transfer with the intention of supporting innovation, providing key information to public policy makers, or dissemination of knowledge from one field of science to another.

Instrumentation/Infrastructure/Data/Software Opportunities:

MRI - Major Research Instrumentation.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5260)

The Major Research Instrumentation Program (MRI) serves to increase access to shared scientific and engineering instruments for research and research training in our Nation's institutions of higher education, and not-for-profit museums, science centers and scientific/engineering research organizations. This program especially seeks to improve the quality and expand the scope of research and research training in science and engineering, by supporting proposals for shared instrumentation that fosters the integration of research and education in research-intensive learning environments. Each MRI proposal may request support for the acquisition (Track 1) or development (Track 2) of a single research instrument for shared inter- and/or intra-organizational use; development efforts that leverage the strengths of private sector partners to build instrument development capacity at MRI submission-eligible organizations are encouraged.

DIBBS – Data Infrastructure Building Blocks.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776)

NSF's vision for a Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) considers an integrated, scalable, and sustainable cyberinfrastructure as crucial for innovation in science and engineering (see www.nsf.gov/cif21). The Data Infrastructure Building Blocks (DIBBs) program is an integral part of CIF21. The DIBBs program encourages development of robust and shared data-centric cyberinfrastructure capabilities to accelerate interdisciplinary and collaborative research in areas of inquiry stimulated by data.

Effective solutions will bring together cyberinfrastructure expertise and domain researchers, to ensure that the resulting cyberinfrastructure components address the researchers' data needs. The activities should address the data challenges arising in a disciplinary or cross-disciplinary context. (Throughout this solicitation, 'community' refers to a group of researchers interested in solving one or more linked scientific questions, while 'domains' and 'disciplines' refer to areas of expertise or application).

This solicitation includes two classes of awards:

- Pilot Demonstration Awards, *** up to $500K/yr for 3 yrs
- Early Implementation Awards. *** up to $1 million/yr for 5 yrs

The Pilot Demonstration projects should address broad community needs of interest either to a large number of researchers within a research domain, or extending beyond that to encompass other disciplines. Early Implementation projects are expected to be of interest to multiple research communities in multiple scientific and engineering domains; these projects will develop frameworks that provide consistency or commonality of design across communities, ensuring that existing conventions and practices are appropriately recognized and integrated, and, most importantly, that the real needs of the community are identified and met.

SSE & SSI (SI2 - SSE&SSI) – Software Infrastructure for Sustained Innovation (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504865&org=CHE&from=home)

Software is an integral enabler of computation, experiment and theory and a primary modality for realizing the Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) vision, as described inhttp://www.nsf.gov/pubs/2010/nsf10015/nsf10015.jsp. Scientific discovery and innovation are advancing along fundamentally new pathways opened by development of increasingly sophisticated software. Software is also directly responsible for increased scientific productivity and significant enhancement of researchers' capabilities. In order to nurture, accelerate and sustain this critical mode of scientific progress, NSF has established theSoftware Infrastructure for Sustained Innovation (SI2) program, with the overarching goal of transforming innovations in research and education into sustained software resources that are an integral part of the cyberinfrastructure.

SI2 is a long-term investment focused on catalyzing new thinking, paradigms, and practices in developing and using software to understand natural, human, and engineered systems. SI2's intent is to foster a pervasive cyberinfrastructure to help researchers address problems of unprecedented scale, complexity, resolution, and accuracy by integrating computation, data, networking, observations and experiments in novel ways. NSF expects that its SI2 investment will result in robust, reliable, usable and sustainable software infrastructure that is critical to achieving the CIF21 vision and will transform science and engineering while contributing to the education of next generation researchers and creators of future cyberinfrastructure. Education at all levels will play an important role in integrating such a dynamic cyberinfrastructure into the fabric of how science and engineering is performed.

It is expected that SI2 will generate and nurture the interdisciplinary processes required to support the entire software lifecycle, and will successfully integrate software development and support with innovation and research. Furthermore, it will result in the development of sustainable software communities that transcend scientific and geographical boundaries. SI2 envisions vibrant partnerships among academia, government laboratories and industry, including international entities, for the development and stewardship of a sustainable software infrastructure that can enhance productivity and accelerate innovation in science and engineering. The goal of the SI2 program is to create a software ecosystem that includes all levels of the software stack and scales from individual or small groups of software innovators to large hubs of software excellence. The program addresses all aspects of cyberinfrastructure, from embedded sensor systems and instruments, to desktops and high-end data and computing systems, to major instruments and facilities. Furthermore, it recognizes that integrated education activities will play a key role in sustaining the cyberinfrastructure over time and in developing a workforce capable of fully realizing its potential in transforming science and engineering.

The SI2 program includes three classes of awards:

1. Scientific Software Elements (SSE): SSE awards target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.

2. Scientific Software Integration (SSI): SSI awards target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems faced by NSF researchers in one or more areas of science and engineering. SSI awards will result in a sustainable community software framework serving a diverse community or communities.

3. Scientific Software Innovation Institutes (S2I2): S2I2 awards will focus on the establishment of long-term hubs of excellence in software infrastructure and technologies, which will serve a research community of substantial size and disciplinary breadth.

Higher Education Opportunities:

REU – Research Experiences for Undergraduates.
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5517&from=fund)

The Research Experiences for Undergraduates (REU) program supports active research participation by undergraduate students in any of the areas of research funded by the National Science Foundation. REU projects involve students in meaningful ways in ongoing research programs or in research projects specifically designed for the REU program. This solicitation features two mechanisms for support of student research: (1) REU Sites are based on independent proposals to initiate and conduct projects that engage a number of students in research. REU Sites may be based in a single discipline or academic department or may offer interdisciplinary or multi-department research opportunities with a coherent intellectual theme. Proposals with an international dimension are welcome. (2)REU Supplements may be included as a component of proposals for new or renewal NSF grants or cooperative agreements or may be requested for ongoing NSF-funded research projects.

NRT – NSF Research Traineeship (replaces IGERT)
(http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505015)

The NSF Research Traineeship (NRT) program is designed to encourage the development of bold, new, potentially transformative, and scalable models for STEM graduate training that ensure that graduate students develop the skills, knowledge, and competencies needed to pursue a range of STEM careers. The NRT program initially has one priority research theme - Data-Enabled Science and Engineering (DESE); in addition, proposals are encouraged on any other crosscutting, interdisciplinary theme. In either case, proposals should identify the alignment of project research themes with national research priorities and the need for innovative approaches to train graduate students in those areas. NRT projects should develop evidence-based, sustainable approaches and practices that substantially improve STEM graduate education for NRT trainees and for STEM graduate students broadly at

an institution. NRT emphasizes the development of competencies for both research and research-related careers. Strategic collaborations with the private sector, non-governmental organizations (NGOs), government agencies, museums, and academic partners that enhance research quality and impacts and that facilitate development of technical and transferrable professional skills are encouraged. Creation of sustainable programmatic capacity at institutions is an expected outcome. Proposals accordingly are expected to describe how institutions will support the continuation and institutional-level scaling of effective training elements after award closure.

Appendix 5

Biographics Sketches of the Authors

Donna G. Blackmond received a Ph.D. in Chemical Engineering from Carnegie-Mellon University in 1984. She has held professorships in chemistry and in chemical engineering in the US, Germany, and the UK, and she has worked in industrial research in the pharmaceutical industry at Merck & Co., Inc. In 2010 she moved from a research chair and joint professorial appointments in chemistry and chemical engineering at Imperial College London to her present position as Professor of Chemistry at The Scripps Research Institute in La Jolla, California.

Professor Blackmond has received Royal Society of Chemistry awards in Physical Organic Chemistry and in Process Technology, a Royal Society Wolfson Research Merit Award and an ACS Arthur C. Cope Scholar Award. She has been a Woodward Visiting Scholar at Harvard and a Miller Institute Research Fellow at Berkeley. She received the Max-Planck-Society's Award for Outstanding Women Scientists and she was an NSF Presidential Young Investigator. She has received the Paul H. Emmett Award in Fundamental Catalysis from the North American Catalysis Society and the Paul Rylander Award from the Organic Reactions Catalysis Society. In 2013, Professor Blackmond was elected as a member of the US National Academy of Engineering.

Professor Blackmond's research focuses on kinetic and mechanistic studies of catalytic reactions for pharmaceutical applications, including asymmetric catalysis. She has pioneered the development of Reaction Progress Kinetic Analysis (RPKA), which makes use of in-situ tools to monitor reaction progress and employs novel graphical manipulations for rapid and straightforward analysis of the kinetics of solution-phase reactions.

Other major areas of Prof. Blackmond's research include the investigation of nonlinear effects of catalyst enantiopurity in stoichiometric, catalytic and autocatalytic reactions as well as studies of enantioenrichment based on the phase behavior of chiral molecules. This work has led both to practical application as well as to fundamental studies probing the origin of the homochirality of biological molecules. She was invited by the Royal Swedish Academy of Sciences to speak at a Nobel Workshop "On the Origin of Life" in Stockholm, 2006. In 2013 she was named a Simons Investigator for fundamental investigations of the origin of life.

Nick Thomson graduated from the University of Edinburgh with a BSc (Hons) in Environmental Chemistry before completing a PhD in organic synthesis under Prof. Gerry Pattenden at the University of Nottingham, England. Nick worked briefly at Zeneca FCMO, Grangemouth (UK) and joined Pfizer, Sandwich (UK) in 1997 as a synthetic chemist in Process Research & Development. Nick spent his early Pfizer career in the evolving process chemistry departments in Sandwich (UK), Sittingbourne (UK) and Holland, Michigan (USA). From 2005 to 2010, Nick led the Sandwich Research Active Pharmaceutical Ingredient (API) department, with accountability for delivery of API technology from lead development to proof of concept. In 2011, Nick joined the Pfizer Chemical Research and Development department in Groton, Connecticut (USA), as a Director of API Process Chemistry Laboratories, with accountability for the Quality by Design development and submission of late stage candidates. In 2014, Nick became head of the Technology API line for Pfizer Chemical Research and Development, with accountability for API Chemical Technologies, Biocatalysis and Computational Chemistry. Nick chairs the International Consortium for Innovation and Quality in Pharmaceutical Development (IQ) API Leadership Group and Working Group on Quality by Design.

Cover design by Janet Hightower
BioMedical Graphics
The Scripps Research Institute, La Jolla, CA