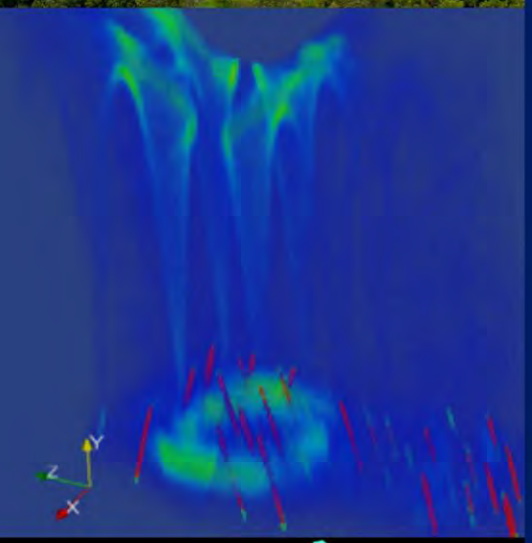

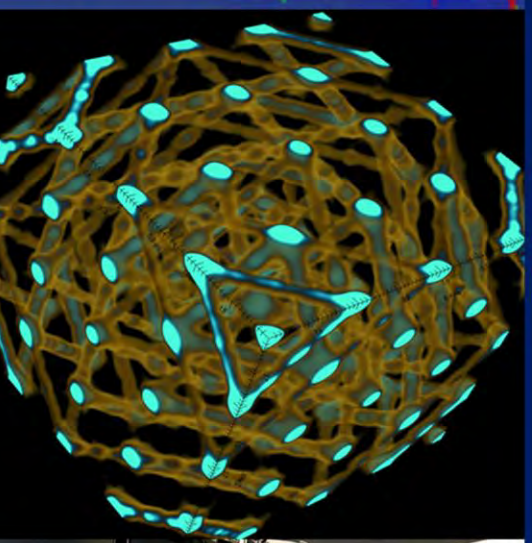

FRONTIERS IN DATA, MODELING, AND SIMULATION

Workshop Report
Argonne National Laboratory
March 30-31, 2015

Organizers:
Peter Littlewood (Argonne National Laboratory)
Thomas Proffen (Oak Ridge National Laboratory)

Sponsored by:
Oak Ridge National Laboratory

FRONTIERS IN DATA, MODELING, AND SIMULATION

Workshop Report

Argonne National Laboratory, March 30–31, 2015

Organizers: Peter Littlewood (Argonne National Laboratory) and

Thomas Proffen (Oak Ridge National Laboratory)

Sponsored by:

Oak Ridge National Laboratory

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

In the past decade or so, neutron and light source user facilities have seen a tremendous increase in experimental capabilities and detector coverage, and as a result, researchers are able to collect ever-larger amounts of data more rapidly. At the same time, increasing computational capabilities are driving the ability to model material structure and dynamics and related properties for larger and more complex systems. Neutron scattering in particular enables simultaneous measurement of structural and dynamic properties of materials from the atomic scale (0.1 nm, 0.1 ps) to the mesoscale (1 μm, 1 μs), matching current capabilities of computational modeling, and the simplicity of the scattering cross section allows the straightforward prediction of related neutron-scattering data from materials simulations.

However, reaching the ultimate goal to accelerate discovery by efficiently utilizing experimental and computational capabilities by the wider science community requires overcoming a number of hurdles. This report summarizes the workshop entitled *Frontiers in Data, Modeling, and Simulation*, which brought together experts in scattering science, theory, applied mathematics, and computational science to discuss the opportunities provided by advances in experiments and computing and to make recommendations for moving toward fully exploiting the scientific potential of these advances.

The body of the report goes into detail about the different science- and technique-based sessions. In summary, we make the following observations and recommendations based on the workshop presentations and discussions, taking into account the recommendations and findings from the earlier grand challenge workshops summarized in the next section.

- **Advances in theory**—non-equilibrium and active systems, larger simulations to address both size and time scales, disordered and heterogeneous systems with numerous interfaces.
- **Improvements in advanced visualization and data analytics tools**.
- **Open access to neutron scattering data** in a form allowing theory validation—ensuring curation of data and proper description of uncertainties.
- **A computing ecosystem that allows easy scaling of materials simulations** as needed.
- **Investment in infrastructure** for data sharing, analytics, and propagation of standards for the materials science community. An operating paradigm is needed for developing, maintaining software platforms—as well as automated processes for archiving and curating data generated by a large and diverse community of experimenters, theoreticians, and modelers.
- **An Innovative environment for theory, algorithms, simulations**, and mechanism to transition promising approaches and prototypes into usable applications for a wider community.
- **Real partnerships between computer scientists, experimental scientists, theorists, and applied mathematicians**.
- **Elimination of the language barrier** between domain scientists and computer scientists and applied mathematicians through joined workshops, training, and workforce development.

II. INTRODUCTION

This report summarizes the findings of the Frontiers in Data, Modeling, and Simulation Workshop, held at Argonne National Laboratory on March 30 and 31, 2015. During the past two years, the Neutron Sciences Directorate at Oak Ridge National Laboratory has sponsored four other grand challenge workshops aiming to create a strategy and a science plan for neutron scattering for the next ten years:

- Grand Challenges in Quantum Condensed Matter, held at The University of California (UC), Berkeley;
- Grand Challenges in Biological Neutron Scattering, held at UC Davis;
- Grand Challenges in Soft Matter, held at UC Davis; and
- Frontiers in Materials Discovery, Characterization, and Application, held in Schaumburg, Illinois.

A summary of the relevant recommendations of the four prior grand challenge workshops is given in the next section.

One emerging theme in the reports of all four workshops is the importance of high-performance computing (HPC), modeling, and simulation to enabling high-impact science. Neutron scattering intensities can be calculated straightforwardly from materials models, making the neutron a unique probe for model validation. It became clear that an additional overarching grand challenge workshop focusing on data, modeling, and simulation would be needed.

The Frontiers in Data, Modeling, and Simulation Workshop addressed that gap. Participants were asked to help in defining the future needs in data, modeling, and simulation as relevant to all aspects of neutron sciences. The workshop was chaired by Peter Littlewood, from Argonne National Laboratory, and was co-organized by Thomas Proffen and Alan Tennant, both from the Neutron Sciences Directorate at Oak Ridge National Laboratory. A total of 29 participants from academia and national laboratories attended the workshop, bringing together expertise in quantum materials, soft matter, biology, applied mathematics, and computer science.

The workshop featured five focus areas: Hard and Quantum Materials; Soft Materials; Materials by Design; Bio Materials and Data, Math, Methods, Analysis and Computing. Each focus area was introduced by a keynote speaker. Each introduction was followed by a number of short presentations and an in-depth discussion. On the second day, Colin Broholm, from Johns Hopkins University, gave the opening keynote lecture, entitled “High-Performance Computing, a Neutron Scatterer’s Perspective.”

This report is structured following the five focus areas. Each section summarizes the talks and identifies challenges and opportunities discussed in each session. The workshop agenda with titles of the talks is included in Appendix I, and a list of workshop participants can be found in Appendix II. The original charge letter can be found in Appendix III. (The original workshop title, “Grand Challenges for Neutrons and Supercomputing,” was subsequently changed to “Frontiers in Data, Modeling, and Simulation” to match the broader scope of the workshop.)

RECOMMENDATIONS FROM PRIOR GRAND CHALLENGE WORKSHOPS

In this section, we briefly list the recommendations related to data, modeling, and simulation given reports for the four earlier grand challenge workshops.

Quantum and Condensed Matter: Better coupling of neutron scattering with theory and simulation: The weak scattering nature of neutron scattering results in a cross-section for both elastic and inelastic scattering which is well understood and can be exactly calculated by many theory and simulation techniques. Close coupling between theoretical efforts and neutron scattering is necessary to make progress in forefront problems. For neutron scattering, this requires detailed understanding of the instrumental resolution function, measurements in absolute units, and development of tools to enable quantitative comparison of theory and experiment.

Biological Neutron Scattering: Develop improved computational methods and tools that exploit high performance computing and that can integrate several/diverse experimental techniques with models and calculations. Data sharing: Develop integrated repositories for sharing data and computational tools for seamless access to complementary data for model building and systems analysis.

Soft Matter: Expand capabilities for computationally-involved data analysis a. Integrate with data simulation methods that can be constrained using data and supplementary information; b. Develop tools for more detailed computation/simulation methods where complex data can be modeled in real time for data where "traditional" analysis fails; c. Real time data visualization; d. Dedicated high speed computers will be required; e. Theory combined with experiment – key component to all these challenges is harmonizing both approaches.

Frontiers in Materials Discovery, Characterization and Application: Chemical Spectroscopy: High Performance Computing should be vigorously pursued in conjunction with Neutron Chemical Spectroscopy; Create software tools that open access to new communities and disseminate understanding of the technique; Libraries and databases are needed to provide reference spectra and models for future science. Materials Science and Engineering: Live data analysis that feeds directly back to data acquisition: adoption of "expert" systems to optimize experimental setup and data analysis.

III. SESSIONS

HARD AND QUANTUM MATERIALS

The opening plenary talk, entitled "Coupling Theory and Numerical Simulations with Spectroscopies in Quantum Materials" was given by **Tom Devereaux,** from the SLAC National Accelerator Laboratory. Tom pointed out that the field of structural biology has become (more or less) fully automated and that the development was driven by dramatic improvements in experimental throughput as well as theory and experiment having evolved to a mature state. Increased funding a decade or so ago and broad involvement from national laboratories, academia, and industry have enabled this transformation in structural biology. The talk focused on the frontier of dynamics and the tremendous opportunities unlocked by the availability of third- and fourth-generation light sources as well as advanced neutron sources. Excitations and dynamics underlie most of the modern grand challenges. A particular limitation to progress in understanding of dynamics is the fact that most theory is focused on the ground state of matter. As a result, theory and modeling to understand state-of-the-art pump probe experiments of materials in excited states are rudimentary and struggling to keep pace. Additional challenges are the often huge parameter space, degrees of freedom of the system, many-particle response functions, and out-of-equilibrium dynamics. On the other hand, a large number of numerical approaches exist to model many of these systems. A schematic overview is shown in Figure 1.

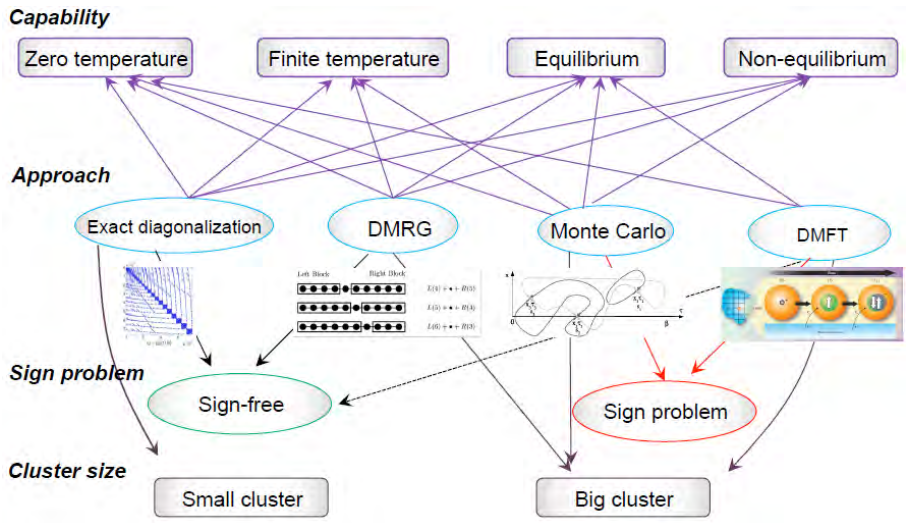

Figure 1. Schematic of numerical approaches.

Tom pointed out that experimental facilities have tremendously progressed but that theory in many ways is still at its first generation. What is needed is a concentrated effort to move theory forward. One such project is the European Theoretical Spectroscopy Facility (ETSF, http://www.etsf.eu/). We need similar efforts, especially in light of ever-improving experimental facilities producing data at a higher rate, ultimately allowing us to systematically sweep phase diagram space and extract large amounts of spectroscopy data.

The main talk was followed by three short presentations. **Maria Fernandez-Serra,** from Stonybrook University, presented progress and remaining challenges related to first-principles calculations of liquid

water. She discussed the importance of incorporating experimental information directly in our modeling methods, in particular focusing on liquid water. Liquid water is a simple system, but from both experiments and simulations, there are a large number of open questions yet to be clarified regarding its atomistic structure. We want to be able to reproduce neutron and x-ray diffraction dynamic and structural data, but when it comes to quantitative comparisons, theorists are unaware of how to incorporate the experimental errors into their numerically obtained dynamical structure factors. These uncertainties depend both on the experiment and on the system under study. She continued to discuss how to introduce this experimental uncertainty in our simulations and how to improve our models using experimental information. A Bayesian approach was used to incorporate a priory or experimental data in our fitting algorithms while maintaining ab initio information. She pointed out that it is important for members of the simulation community to provide uncertainty quantification of their ab initio data.

Gabriel Kotliar, from Rutgers University, gave a talk on strongly correlated materials and theoretical spectroscopy and materials design. The interaction between theory and experiment allows one to design materials with the ultimate goal of being able to make predictions based on a given theory. However, in many cases there is still a long way to go. As an example, in the case of plutonium, spin density functional theory fails because it predicts large orbital and spin moments that are not found experimentally [1]. Dynamical mean field theory (DMFT) is able to describe the observed spectroscopic data. The key to success seems to be using the various tools and methods in combination because together they provide a broader platform for accurate and faster theoretical spectroscopy and material discovery and design. What is needed is the development of a strategy how to apply such a combined approach more routinely.

The last short presentation in the hard materials session was given by **Thomas Maier,** from Oak Ridge National Laboratory, on progress and challenges in cluster DMFT. Understanding, predicting, and controlling the remarkable properties of correlated electron materials remains a grand challenge and materials theory and modeling are essential elements in solving this problem. The challenge for simulations of quantum many-body problems needed to describe these systems arises from the exponential increase of the complexity with system size (number of atoms, orbitals, etc.). Cluster dynamical mean-field theories (CDMFTs) [2], such as the cellular DMFT [3] and the dynamic cluster approximation (DCA) [4], reduce this complexity by mapping the problem onto a finite-size cluster of atoms embedded in a dynamic mean-field host that is designed to represent the remaining degrees of freedom. The resulting cluster problem is then tractable with methods such as quantum Monte Carlo (QMC) [5]. CDMFT techniques have been used extensively to simulate the single-band Hubbard model of the cuprate high-temperature superconductors in order to understand their superconducting, antiferromagnetic, and pseudogap behaviors [2]. An important focus of these studies is the role of spin-fluctuations, which are measured in neutron scattering experiments and which are thought to dominate many aspects of the physics of these systems [6,7].

Recent progress in the QMC algorithms [8] and the DCA framework [9] has enabled more advanced simulations addressing new questions that were out of reach before. Examples are the interplay between the pseudogap behavior and superconductivity [10] and the reliable determination of the

superconducting transition temperature through cluster size scaling [11]. Applying these methods to other correlated electron systems, such as the iron-based superconductors, requires more complex simulations of multi-orbital extensions of the single-orbital Hubbard model. While this remains a highly challenging task, primarily because of the negative fermion sign problem [12], new developments in the QMC sampling schemes [13] and the underlying DCA framework [9] demonstrate that significant progress can be made, putting such simulations on the near horizon.

DISCUSSION

Initial discussion focused on the state of spectroscopy calculation codes. Most of the commonly used codes are being developed in Europe, which has secured the lead by yearlong investments in these tools. They also successfully collaborate on the theory side through initiatives such as the ETSF. The state of the community in the United States was characterized as being more fractured and was described as a cottage industry. Participants agreed that a targeted investment and a much more collaborative approach, particularly between the national laboratories and universities, are needed in the United States. It was argued that theory groups, which currently tend to work in isolation, may need to start providing support to code developers.

The second main discussion topic involved comparisons between materials simulations, theoretical predictions, and experimental data. Quoting Simon Billinge, “The method of validating theory against experiment by comparing one paper to another is a 19th century way of working. Communities need a more integrated approach where both modeling and experimental data are analyzed together.” Two main hurdles were identified: Discovery and open access to relevant experimental and simulation data and the relevant tools to allow calculation of the properly corrected experimental quantities from a theoretical model or simulation. Clearly the national user facilities are playing a critical role addressing those points. The second issue discussed was basically about trusting the data and providing proper uncertainties and details (e.g., about the sample, instruments) to make the data meaningful to the wider scientific community. A related issue briefly touched in the discussion was the need to give proper credit when using open data.

In summary the common themes and recommendations from the hard and quantum materials session are:

- Targeted investments in developing relevant theory are needed, and broad collaborations are encouraged.
- Easy and open access to validated experimental and simulation data is needed.
- Proper uncertainty quantification is needed to allow meaningful comparisons of experimental data and simulation results.
- Codes need to be user friendly enough for the nonexpert, and training is needed to teach the next generation about methods, codes, and limitations.

REFERENCES

[1] J. C. Lashley, A. Lawson, R. J. McQueeney, and G. H. Lander, Absence of magnetic moments in plutonium, *Phys. Rev. B* **72**, 054416 (2005)
[2] T. Maier, M. Jarrell, T. Pruschke, M. Hettler, Quantum cluster theories. *Rev. Mod. Phys.* **77**, 1027–1080 (2005).
[3] G. Kotliar, S. Savrasov, G. Pálsson, G. Biroli, Cellular Dynamical Mean Field Approach to Strongly Correlated Systems. *Phys. Rev. Lett.* **87**, 186401–186401 (2001).
[4] M. H. Hettler, A. N. Tahvildar-Zadeh, M. Jarrell, T. Pruschke, H. R. Krishnamurthy, Nonlocal dynamical correlations of strongly interacting electron systems. *Phys. Rev. B*. **58**, R7475–R7479 (1998).
[5] M. Jarrell, T. Maier, C. Huscroft, S. Moukouri, Quantum Monte Carlo algorithm for nonlocal corrections to the dynamical mean-field approximation. *Phys. Rev. B*. **64**, 195130 (2001).
[6] T. A. Maier, M. S. Jarrell, D. J. Scalapino, Structure of the Pairing Interaction in the Two-Dimensional Hubbard Model. *Phys. Rev. Lett.* **96**, 047005–047004 (2006).
[7] T. A. Maier, D. Poilblanc, D. J. Scalapino, Dynamics of the Pairing Interaction in the Hubbard and t-J Models of High-Temperature Superconductors. *Phys. Rev. Lett.* **100**, 237001–237004 (2008).
[8] E. Gull *et al.*, Continuous-time Monte Carlo methods for quantum impurity models. *Rev. Mod. Phys.* **83**, 349–404 (2011).
[9] P. Staar, T. Maier, T. C. Schulthess, Dynamical cluster approximation with continuous lattice self-energy. *Phys. Rev. B*. **88**, 115101 (2013).
[10] E. Gull, O. Parcollet, A. J. Millis, Superconductivity and the Pseudogap in the Two-Dimensional Hubbard Model. *Phys. Rev. Lett.* **110**, 216405 (2013).
[11] P. Staar, T. Maier, T. C. Schulthess, Two-particle correlations in a dynamic cluster approximation with continuous momentum dependence: Superconductivity in the two-dimensional Hubbard model. *Phys. Rev. B*. **89**, 195133 (2014).
[12] M. Troyer, U. Wiese, Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations. *Phys. Rev. Lett.* **94**, 170201–170201 (2005).
[13] Y. Nomura, S. Sakai, R. Arita, Multiorbital cluster dynamical mean-field theory with an improved continuous-time quantum Monte Carlo algorithm. *Phys. Rev. B*. **89**, 195146 (2014).

SOFT MATERIALS

Soft matter is a class of materials that is referred to as “soft” because their structures are typically easily deformed by temperature, composition, external stimuli, or other associated variables. This sensitivity arises from the competition among different interactions and between the interactions and entropy. Soft materials are also highly correlated many-body systems, with complex structures, dynamics, and transport processes spanning a wide range of length and time scales. Many soft materials are disordered, so the typical tools of solid-state physics cannot be easily applied. Soft materials may also be far from equilibrium, so they cannot be understood using standard statistical mechanics. Soft materials are also distinguished from traditional hard materials by the multiplicity of time and length scales involved in processes governing their behavior and by their degree of disorder (e.g., mostly defects with a few crystals as opposed to mostly crystalline with a few defects). As such, there are many challenges in interfacing disparate methods appropriate at different scales and in managing the massive data sets arising from tandem use of multiscale experiments and simulations (including high-throughput

approaches), including the establishment of standards for data collection, data mining and analysis, and sustainable storage.

The opening talk in the soft materials session was given by **Monica Olvera de la Cruz,** from Northwestern University, who reminded participants that soft materials constitute the basic component of living systems, are integrated into the fabric of modern society, and will play a key role in futuristic devices. Examples include fibers, membranes, tissues, polyelectrolytes, gels, biomimetic systems, ionic-liquids, and ion-containing polymers. Nearly everything that is disordered and described by classical Hamiltonians or Lagrangians is part of soft matter. Soft matter problems that are part of the Materials Genome Initiative, and opportunities in the field of soft matter that lie at the interfaces with other fields such as biology and energy production, storage, and conversion were identified. Two topics were addressed: (1) Materials by Design and (2) Bridging Space and Time Scales. Within the Materials by Design topic, the goal is to attack the inverse problem. That is, starting with a function or property, work proceeds backwards to design the material constituents and the process to produce it. Designing polymers, liquid crystals, gels, and supramolecular structures requires experimental data. The question is how to incorporate the experiments with theory and modeling. That requires the development of frameworks to link molecular composition to geometry, continuum elasticity, electrostatics, and statistical mechanics of meso-ordered matter. Understanding the principles that link mesoscale ordering (such as programmable shape and function in reconfigurable materials) will advance materials synthesis and manufacturing (i.e., 3D printing). Another subject that required attention in soft materials is control of charge transport across membranes. The grand challenges include describing systems with nonlocal interactions and non-equilibrium phenomena. In the topic of Bridging Space and Time Scales, recent computational tools have led to major discoveries in soft material. The challenges are validation of the larger-scale methods from modeling of finer-scale methods; the opportunities are new developments in advanced computer architectures and software. The outcome is the potential to overlap time and length scales to cohesively investigate multiple scales.

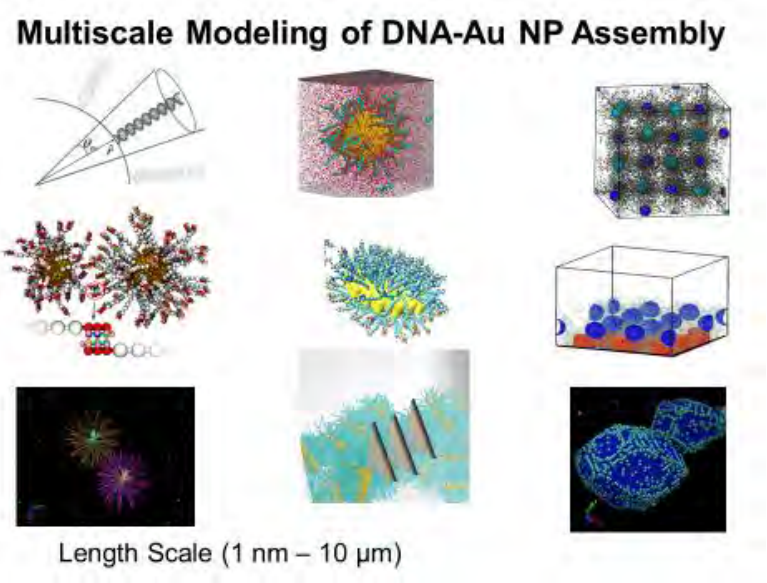

Figure 2. Multiscale modeling of DNA-Au nanoparticle assembly.

Multiscale modeling of DNA-functionalized gold nanoparticles was described (Figure 2). The question on how to create single crystals using DNA functionalized nanoparticles (NPs) was addressed by a combination of models and techniques, including molecular

dynamics (MD) simulations with explicit chains and hybridization but implicit solvent, colloidal (short-range attraction + Yukawa repulsion) MD models, classical density functional theory, and MD studies with explicit chains and ions (implicit water). MD and Monte Carlo simulations were used to describe layer-by-layer epitaxial growth of DNA-functionalized NPs on DNA-modified patterned substrates and to engineering defects. The MD scale-accurate model with explicit chains and hybridization reproduces the experimentally reported crystalline structures in binary DNA-functionalized NPs and demonstrates that active hybridization is crucial to achieve crystallization. The construction of the Wulff shape of single crystal by computing the surface energies of the crystalline structure along different planes is achieved by using this scale accurate model. On the other hand, the colloidal model (with parameters obtained from the scale-accurate model and classical density functional theory) describes the kinetics of crystallization. Classical density functional theory and MD results with implicit chains and ions combined with experimental results show that the interactions among NPs at high salt concentrations (around 1 M NaCl or 50mM $CaCl_2$) are long range. In the area of layer-by-layer growth, it is found that DNA hybridization between complementary DNA linkers on nanoparticles and on a substrate can precisely lead the DNA-coated nanoparticles to desired template sites and to form a defect-free epitaxial layer as well as to engineering new structures of "ordered" defects that resemble liquid-crystalline phases of NPs.

The next talk, entitled "Trivalent ions change the sign of interactions between polymer brushes," was given by **Matt Tirrell,** from the University of Chicago. Applications of end-tethered polyelectrolyte "brushes" to modify solid surfaces have been developed and studied for their colloidal stabilization and high lubrication properties. Current efforts have expanded into biological realms and stimuli-responsive materials. His group explores responsive and reversible aspects of polyelectrolyte brush behavior when polyelectrolyte chains interact with oppositely charged multivalent ions and complexes as counterions. There is a significant void in the polyelectrolyte literature regarding interactions with multivalent species. Our work demonstrates that interactions between solid surfaces bearing polyelectrolyte brushes are highly sensitive to the presence of trivalent lanthanum, La^{3+}. Lanthanum cations have unique interactions with polyelectrolyte chains, in part due to their small size and hydration radius, resulting in a high local charge density. Using La^{3+} in conjunction with the surface forces apparatus, adhesion has been observed to reversibly appear and disappear upon the uptake and release, respectively, of these multivalent counterions. In media of fixed ionic strength set by monovalent sodium salt, at I_0 = 0.003 M and I_0 = 0.3 M, the sign of the interaction forces between overlapping brushes changes from repulsive to attractive when La^{3+} concentrations reach 0.1 mole % of the total ion concentration. These results are also shown to be generally consistent with, but subtlety different from, previous polyelectrolyte brush experiments using trivalent ruthenium hexamine as the multivalent counterion. There is no well-founded theory that predicts this behavior in multivalent ions.

The final presentation in this session was given by **Bobby Sumpter,** from Oak Ridge National Laboratory, who discussed integrating multimodal data (multiple measurements) with simulations in order to create a discovery and innovation ecosystem for materials science. This requires tight integration between modeling and simulations, measurement, data analysis, theory, and synthesis [1]. We need to move

toward a philosophy of “Compute not for numbers but to procure understanding” (W. Kohn). Today’s materials require the ability to operate at extreme use and under extreme conditions; they need robust properties over long lifetimes even under extreme environments. Soft matter is often characterized by complex free energy landscapes with multiple metastable minima [2]. A pervasive problem is that the systems can become kinetically trapped in metastable minima, unable to reach the equilibrium state. On the other hand, some of these minima might have desirable structures and properties. A key challenge is to understand how to avoid some kinetic traps and to exploit others to optimize function.

Current issues in modeling and simulations include the ability to carry out large-scale MD simulation for the appropriate length and time scales. This is particularly daunting in the study of hierarchical assemblies. Other issues are in regard to the lack of accurate force fields that can account for charge transport, polarizability, and chemical reactions in a trustworthy fashion and to a lack of systematic approaches to coarse-graining potentials. We also need to move toward some standardized and validated methods to couple models at interfaces between different computational zones. Improved methods are needed to address computational inefficiencies associated with the time-sampling requirements of the fastest components. Also, new techniques are needed to address difficult materials phenomena that engage all length and time scales simultaneously, such as thermal transport of electronic degrees of freedom. Kohn also pointed to the area of non-equilibrium processes, which is in need of considerable development [3].

DISCUSSION

Discussions focused on the urgent need to expand capabilities for computationally involved data analysis. The following recommendations were made:

- Integrate with data simulation methods that can be constrained using data and supplementary information.
- Develop tools for more detailed computation/simulation methods where complex data can be modeled in real time for data where “traditional” analysis fails.
- Real time data visualization and data analytics on the time frame of an experiment.
- Obtain the required dedicated high-speed computers.
- Combine theory with experiment—A key component to all these challenges is harmonizing both approaches.
- Meet the desperate need for an infrastructure for data sharing, analytics, and propagation of standards for the soft matter community. At present, there is redundancy of effort and lost opportunities for data use, analytics, and mining. An operating paradigm is needed for developing and maintaining software platforms and automated processes for archiving and curating data generated by the large and diverse community of experimenters, theoreticians, and modelers.
- Make progress toward developing a deep understanding of how the dynamic response of active matter can be controlled by the amplification of molecular-level stimuli-responsive materials.

- Develop a framework for understanding information out of equilibrium.
- Work on viable ways to treat the “inverse problem” [i.e., working backward from a desired set of materials properties (e.g., composition, molecular weight) to design a system with those properties. One place with a big impact is to design a process that could involve both equilibrium and non-equilibrium steps and that leads to a material with the desired set of properties.

In bio-based materials, HPC has helped extend length scales but not time scales. There are lots of ad hoc solutions, but effectively extending time scales would take some careful thought and corroboration by experimental validation.

REFERENCES

[1] S. Kalinin, B. G. Sumpter, R. Archibald, Big-Deep-Smart Data in Imaging for Guiding Materials Design, *Nature Materia*ls, **14**, 973-980 (2015).
[2] Stephen Z. D. Cheng, Andrew Keller, The Role of Metastable States in Polymer Phase Transitions: Concepts, Principles, and Experimental Observations, Annual Review of Materials Science 28, 533-562 (1998).
[3] Non-equilibrium Phenomena in Confined Soft Matter, Springer (2015) ISBN 978-3-319-21948-6. Editor, Simone Napolitano

BIO MATERIALS

The session on bio materials was opened by **Benoit Roux,** from the University of Chicago, discussing how neutron science could benefit the scientific community in the area of biomolecular systems and could help identify the science grand challenges for that community. MD simulations of large biological macromolecules have reached the point where they can be used to provide meaningful insight on the function of complex systems. The results from those simulations raise intriguing questions about the participation of water molecules that may be addressed using neutron scattering. This was illustrated with two recent examples on K+ channels, and the sodium-potassium (Na/K) ATPase pump.

Activation of a K+ channel typically leads to a transient period of ion conduction until the selectivity filter spontaneously undergoes a conformational change toward a constricted nonconductive state (inactivation). Subsequent removal of the stimulus closes the gate and allows the selectivity filter to return to its conductive conformation (recovery). The recovery process can take up to several seconds, an extraordinarily long time. Yet the structural differences between the conductive and inactivated filter are very small. MD simulations revealed that structural water molecules bound directly behind the selectivity filter are directly responsible for the long time for the recovery from inactivation. This prediction from MD was verified with functional experiments [1] and by nuclear magnetic resonance (NMR) [2].

The Na/K pump is an ATPase that generates Na+ and K+ concentration gradients across the cell membrane. For each ATP molecule, the pump extrudes three Na+ and imports two K+ by alternating between outward- and inward-facing conformations that preferentially bind K+ or Na+, respectively. Remarkably, the selective K+ and Na+ binding sites share several residues, and how the pump is able to achieve the selectivity required for the functional cycle is unclear. Free-energy MD simulations reveal that protonation of the acidic side chains involved in the binding sites is critical to achieve the proper K+ selectivity [3,4]. The gating charge detected upon ion dehydration and binding to the pump has been calculated and correlated with experiments. [5]

The main talk was followed by two shorter presentations. The first one was given by **Loukas Petridis,** from Oak Ridge National Laboratory, entitled "Addressing Challenges in Biology: Combining Neutrons and HPC." A central dogma in biology is the so called "structure-function" relationship, that the three-dimensional structures of biomolecules define their biological function. In the past decades, structural biology methods, such as x-ray crystallography (XRC) and NMR, have been successfully applied to determine the structure of single-domain, well-folded proteins. Conformational flexibility, which is manifested as transitions among multiple states accessible to a macromolecular complex, is of acute importance to the function of biomolecular systems. Understanding conformational flexibility across multiple spatial and temporal scales is still at a primitive stage. For example, crystallizing flexible proteins for XRC has proven to be particularly challenging, and applications of NMR to large complexes is challenging due to the difficulty in assigning NMR correlations. This limitation has been particularly noticeable in structural and dynamic studies of large-scale flexible systems, such as kinases and intrinsically disordered proteins.

Small angle neutron scattering (SANS) is an ideal technique to provide information on the relative arrangement of components within functioning complexes, in particular, when they are disordered. Most of the biological systems characterized using neutrons are of a complexity such that the direct interpretation of experiment with analytical theory cannot be made unequivocally. Numerical simulation is therefore required. In particular, neutron scattering probes time- and length-scales that are similar to those of MD simulation. Hence, MD has become invaluable in the interpretation of neutron data. The synergy of neutron scattering experiment and MD simulation is achieved by performing simulations of the same systems experimented upon under the same environmental conditions (e.g., solvent, temperature, pressure). Simulation and experiment are then bridged by calculating relevant neutron scattering quantities [e.g., I(Q), S(Q,ω), I(Q,t).] directly from the simulation results and comparing with experimental results.

Atomistic MD simulation on supercomputing platforms now scales up to O(100k) cores, permitting time scales of ~1 μs and ~10^7 atom system sizes. However, the future of molecular-scale supercomputing is likely to be oriented to ensemble-based methods, which involve performing multiple simulations of the same, smaller system in a way that improves the sampling of configurational space and takes full advantage of massively parallel supercomputers.

Generation of SANS-consistent ensembles. The relative populations of different system states determined using the above simulation methodologies can be refined by optimizing the theoretical scattering profiles against experimental SANS data. A challenge will be to help address the common over-fitting problem in ensemble-based SANS refinement. These ensemble optimization methods are based on the computational generation of conformations, computation of a SANS I(Q) profile for each conformation, and subsequent selection of conformations that are consistent with the experimental SANS. These methods face a steep challenge when applied to large flexible macromolecular complexes, in that structurally distinct conformations yield similar I(Q) profiles. This degeneracy (Figure 3), makes it very challenging to unequivocally predict a structural ensemble using current methods. Our approach to overcome this hurdle is to use contrast variation studies of selectively deuterated proteins or segments as a means of providing additional constraints to the simulation. This ensures each conformation has a unique set of I(Q) profiles, performed at different match points (ratios of H_2O to D_2O), that describe it. For example, the first two conformations in Figure 3 have similar I(Q) when in a 100% D_2O solvent and cannot be differentiated. However, when SANS is performed on the same system at a different match point (40% D_2O), these conformations yield different I(Q). Therefore, combining I(Q) from multiple solvents overcomes degeneracy. Moreover, contrast matching with selective deuteration, another distinct advantage of neutron scattering, will provide multiple restraints for each system. Hence, this new approach will bridge a gap between neutron experiment and simulation, providing a molecular-level view of the dynamics of complex biomolecular systems.

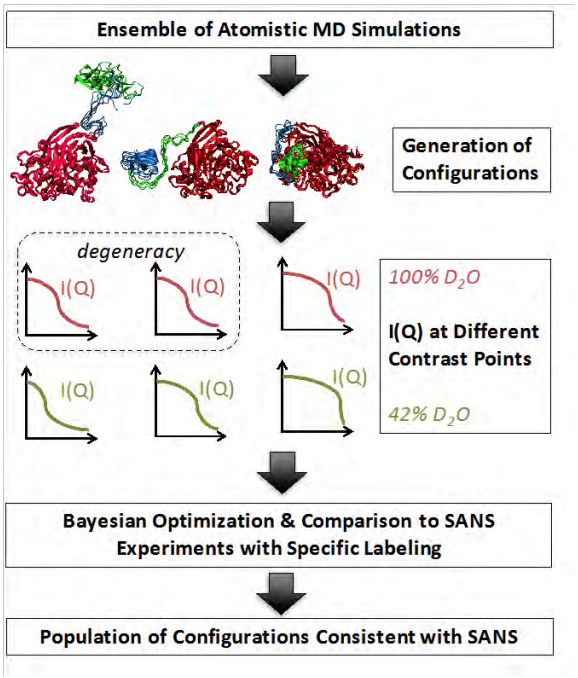

Figure 3. Generation of SANS-consistent ensembles.

The next presentation was given by **Pratul Agarwal,** from Oak Ridge National Laboratory, on Neutrons, Structures and "Models" in Biology. Computing continues to make tremendous impact on biology. In addition to hypothesis generation, design of experiments, collection of data, and interpretation, the big impact of computational methods is coming from development of models. The models that combine data from different experimental techniques and across scales are able to provide vital knowledge about biological processes ranging from processes that involve a single molecule to full ecosystems. The need of the hour, as is widely acknowledged, is the development of high-quality, validated models of molecular and cellular processes that allow the making of testable predictions. Unfortunately, the current models, especially of the molecular processes, remain limited and heavily rely on a single experimental technique. In addition, there are significant delays (months to years) associated with the experimental design, data gathering, and analysis.

Joint computational-experimental investigations that utilize neutron scattering (and related techniques) are poised to make significant impact on our understanding of biomolecular processes. Structural information about an increasing number of proteins (and other biomolecules) is being generated every day. In addition, techniques such as quasi-elastic neutron scattering and spin-echo are able to provide information about the dynamics of biomolecules. Computing resources still continue to double every 18 months, more or less, in line with Moore's law. Experimental information is also being collected at an exponential pace. The close integration of modeling and simulations with experimental techniques that enable the use of computing to build high-quality dynamical models of biomolecular complexes and even full cells will allow us to investigate important aspects at molecular and cellular levels in full atomistic detail. Development of the first model of a full prokaryotic cell (such as *Escherichia coli*) is now within our grasp. Over 80% of proteins associated with such cells are available, and neutron scattering and other techniques continue to provide information about the biological membranes. MD simulations and other techniques have already shown that it is possible to develop models of biomolecular complexes with 100 million atoms. A simple model of full biological cell is expected to compose of 10 to 100 billion atoms. Combining the next generation of instruments for the Spallation Neutron Source (SNS) and exascale computing resources would make these models a reality.

The benefits of development of a fully atomistic level of models of cells and cellular processes will have an unprecedented effect on research in energy, environment, and human health. These models will allow a detailed understanding of the mechanism of carbon sequestration in algae as well as processes associated with conversion of cellulose to fermentable sugars. Health processes associated with cancer and other diseases can also be investigated at new scales, all of which are areas of great relevance to the DOE mission.

DISCUSSION

There is obvious overlap between the needs and discussion in the soft matter and bio materials sections of this report. The main need identified in this session is the development of force fields for MD simulations that are suitable for biological systems and, in particular, that are able to treat ions. Current developments include the use of polarizable force fields, but as one participant put it, "not all force fields are a caricature of reality—like in a political cartoon, you know who the president is, but you could not use the image to measure the length of his nose."

More recently some people in the field have moved to using approaches based on quantum mechanics, but those approaches can be too computationally expensive. It was noted that it could be damaging to the community to spend too much effort on quantum mechanical detail at the expense of less precise results. Efforts in both areas are needed. Neutron scattering was acknowledged as a crucial tool in providing data to be able to validate models.

REFERENCES

[1] J. Ostmeyer, S. Chakrapani, A. C. Pan, E. Perozo & B. Roux. Recovery from Slow Inactivation in K+ Channels Controlled by Water Molecules *Nature* **501**, 121-124, (2013). PMC3799803
[2] M. Weingarth, E. A. van der Cruijsen, J. Ostmeyer, S. Lievestro, B. Roux & M. Baldus. Quantitative Analysis of the Water Occupancy around the Selectivity Filter of a K(+) Channel in Different Gating Modes. *J. Am. Chem. Soc.* **136**, 2000-2007, (2014). PMID: 24410583 [PubMed - in process]
[3] H. Yu, I. M. Ratheal, P. Artigas & B. Roux. Protonation of key acidic residues is critical for the K-selectivity of the Na/K pump. *Nat. Struct. & Mol. Biol.* **18**, 1159-1163, (2011). PMC3190665
[4] I. Ratheal, G. Virgin, H. Yu, B. Roux, C. Gatto & P. Artigas. Selectivity of externally facing ion binding sites in the Na/K pump to alkali metals and organic cations. *Proc. Natl. Acad. Sci. U.S.A.*, (2010). PMC2972997
[5] J. P. Castillo, H. Rui, D. Basilio, A. Das, B. Roux, R. Latorre, F. Bezanilla & M. Holmgren. Mechanism of potassium ion uptake by the Na(+)/K(+)-ATPase. *Nat. Comm.* **6**, 7622, (2015). PMC4515779

MATERIALS BY DESIGN

The session was opened by the talk on materials by design given by **Thomas Schulthess** from the ETH Zürich. He reminded us that new materials, and in particular, new complex materials are to this date mostly discovered serendipitously. Edison tested 3000 materials before settling on a burned sewing thread for his filament. This type of Edisonian development is still common today, and of course the idea behind materials by design is to step beyond the trial-and-error approach. Next he gave an example of ab initio materials by design: giant tunneling magnetoresistance (TMR). Simple theoretical models explained TMR through an amorphous oxide barrier. Later ab-initio simulations predicted a TMR effect through certain barrier oxides (e.g., for Fe|MgO|Fe junctions) that were two order or magnitude larger [1]. By 2004, crystalline giant TMR junctions had been realized experimentally [2,3] and allowed more sensitive hard drive read heads that are suitable of higher storage density. That success involved a tight interaction between theory and experiment.

One of the driving goals in light of increasing computer power is to determine how quickly we can screen materials computationally. In addition to raw computing power, complicated workflows are required to carry out systematic searchers automatically and help us deal with the vast amount of data produced. As a simple example, there are about 150,000 documented compounds. Some very basic properties computed with Density Functional Theory (DFT) based quantum simulations can take ~10 minutes on a powerful desktop. Scaling this to the OLCF's Titan computer with nodes containing 18,600 central processing units per graphics processing unit would allow one to calculate 18,000 structures in about 10 minutes. Efforts like these are under way (e.g., www.materialsproject.org). Looking at the problem from the other end, we can consider how complex we can make the simulations for realistic systems. If we replace the simple DFT in the example above with a calculation of the electronic structure using the linearized augmented plane wave method, the number of structures that can be screened even on a Titan level machine decreases dramatically. The increases coming with the next-generation (exascale) computers will make these approaches feasible, but big investments in the related scientific software will be needed. Figure 4 shows a schematic view of how things are organized: physical model,

mathematical description, algorithm, or implementation description for the imperative code, which is compiled to run on a given computing architecture. Figure 4 also indicates tasks falling into the realm of domain scientists and applied mathematicians on the left and computer engineers on the right.

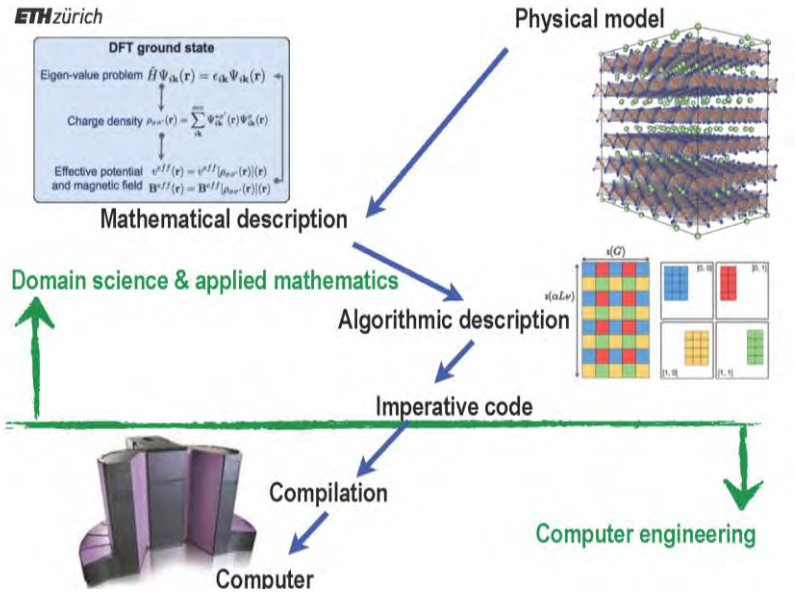

Figure 4. Organization of physical model, mathematical description, and implementation description for the imperative code in the realms of domain scientists and applied mathematicians.

An additional challenge is that we do not have a single computing architecture, and, as we reach the end of Moore's law, the number of architectures optimized for certain applications will only grow. The green line in Figure 4 below the algorithmic description indicates the distinction between the domain scientist and the computer engineer. This transition will require new productivity tools that allow the domain scientists to innovate and try different algorithms. In many ways, what are needed are libraries that are as easy to use as "numpy" to efficiently perform the needed calculations.

Timmy Ramirez-Cuesta from Oak Ridge National Laboratory gave a talk entitled "Neutrons and Numbers: VISION the World's First High-Throughput Inelastic Neutron Scattering Spectrometer." He introduced molecular spectroscopy as a very powerful tool to study the dynamical properties of solids, liquids and gases. Inelastic neutron scattering (INS) is a very powerful tool to study hydrogen-containing materials. With the development of neutron spallation sources and the use of epithermal neutrons, INS can measure the vibrational spectra of materials on the whole range of vibrational motions (0–4400 cm^{-1}) and effectively open up the field of neutron spectroscopy [4]. The recently commissioned VISION spectrometer at the Oak Ridge SNS ,has an increased overall flux at low energy transfers up to 4000 times over its predecessors. INS is a technique that is ideally suited to study hydrogen-containing materials due to the high cross section of hydrogen [4]; it is also the case that INS spectra are straightforward to model [5]. It could even be argued that INS presents the most rigorous experimental test of ab initio methods.

Ramirez-Cuesta described a demonstration of the use of combined INS and DFT to identify the dynamics of the captured molecules trapped within porous materials and presented some examples [7,9]. In the final part of the talk, he discussed the limits of what is possible now with the SNS VISION spectrometer e.g., determining INS spectra of publishable quality in minutes for samples in the gram quantity range [6], or measuring the signal of samples in the milligram range, and the direct determination of the signal

of 2 mmol of $C0_2$ adsorbed on functionalized catalysts. He discussed the challenges that we are facing, in particular, methods to automate data analysis and interpretation through computer modeling.

Sasha Balatsky, from Los Alamos National Laboratory, gave the final talk in the session, entitled "Computation Capability for Design of Complex Electronic Materials." He discussed the existence of various materials databases containing the electronic properties of materials that have been used in the past. He noted that bridging the gap between theory and experiment requires collaboration and co-location of experimentalists and theorists (a common theme during the workshop) as well as the development of theory and simulations that are faithful to the experimental probes.

DISCUSSION

There are currently a number of pressing issues for the general area of what is often referred to as "materials by design." To highlight those, the situation and demand for improved materials should be discussed [10,11]. Briefly, the number of functionalities required for developing and optimizing materials fundamental to our modern technology and infrastructure needs, such as energy storage and capture capacities, energy conversion and transmission efficiencies, robust performance under extreme conditions or extreme use, and the simultaneous delivery of multiple functions (e.g., high strength and low weight) continues to rapidly increase. To meet these demands requires substantially more efficient paradigms for materials discovery and design that go beyond current Edisonian and classical synthesis-characterization-theory approaches or occasional serendipitous discoveries. However, we must recognize that the structures and composition underpinning the next-generation materials are not well understood, much less the pathways to synthesize them. As a prerequisite, this development requires an understanding of materials' hierarchical and heterogeneous structure and dynamics from the atomic scale to real-world components and systems. In addition, understanding and modeling non-equilibrium synthesis and processing are vital to achieving a transformative impact.

At a more refined level, at least **three major gaps** need to be addressed to enable a computationally based framework that facilitates designing materials with desired properties, most importantly including their synthesis pathways. **First**, there is the need for enhanced reliability for the computational techniques in such a way that they can accurately (and rapidly) address the complex functionalities mentioned above, provide the precision necessary for discriminating between closely competing behaviors, and capability of achieving the length/time scales necessary to bridge features such as domain walls, grain boundaries, and gradients in composition. **Second**, there is a need to take full advantage of all of the information contained in experimental data to provide input into computational methods to predict and understand new materials. This includes integrating data efficiently from different characterization techniques to provide a more complete perspective on materials structure and function. **Third**, pathways need to be established for making materials. In general, pathways for making materials are least amenable to theoretical exploration due to the daunting dimensionality (a plethora of metastable states and pathways accessing them) and primarily rely on the expertise of individual researchers. Exascale computing capacities, and big, deep-data approaches offer new

possibilities to bridge the gap. For example, they can be bridged with coarse-grained models to ultimately access mesoscopic scales and to scale up enhanced free energy and kinetic sampling used for the atomistic MD study of complex interfaces. Additionally, big data analytics on existing or evolving bodies of knowledge on synthesis pathways can suggest correlations between materials properties and synthetic routes, potentially providing specific research directions. These two approaches must be integrated and utilized.

REFERENCES

[1] Butler, et al., *Phys. Rev. B* **63**, 54416 (2001)
[2] Parking et al., *Nature Materials* **3**, 862 (2004)
[3] Yuasa et al., *Nature Materials* **3**, 868 (2004)
[4] Mitchell PCH, Parker SF, Ramirez-Cuesta A, Tomkinson J. Vibrational Spectroscopy with Neutrons, with applications in Chemistry, Biology, Materials Science and Catalysis. London: World Scientific; 2005.
[5] AJ Ramirez-Cuesta, MO Jones, WIF David, Materials Today, 12, 2009, 54-61.
[6] Jalarvo, N., Gourdon, O., Ehlers, G., Tyagi, M., Kumar, S. K., Dobbs, K. D., … Crawford, M. K. (2014). The Journal of Physical Chemistry C, 118(10), 5579–5592. doi:10.1021/jp412228r
[7] Yang, S., Sun, J., Ramirez-Cuesta, A. J., Callear, S. K., David, W. I. F., Anderson, D. P., Newby, R., et al. (2012) Nature chemistry, 4(11), 887–94. doi:10.1038/nchem.1457
[8] Yang S, Ramirez-Cuesta AJ, Newby R, Garcia-Sakai V, Manuel P, Callear SK, Campbell SI, Tang CC and Schröder M, (2014) Nature chemistry, doi: 10.1038/nchem.2114
[9] Casco M.E., Silvestre-Albero J., Ramírez-Cuesta A. J., Rey F., Jordá J. L., Bansode A., Urakawa A., Peral I., Martínez-Escandell M., Kaneko K. and Rodríguez-Reinoso F., Nat Commun, vol. 6, Mar. 2015 doi: 10.1038/ncomms7432.
[10] S. Kalinin, B. G. Sumpter, R. Archibald, Big-Deep-Smart Data in Imaging for Guiding Materials Design, *Nature Materials*, **14**, 973-980 (2015)
[11] Bobby G Sumpter, Rama K Vasudevan, Thomas Potok, Sergei V Kalinin, A bridge for accelerating materials by design, *NPJ Comp. Mater*. **1**, 15008 (2015)

COMPUTING, METHODS, AND ANALYSIS

The computing, methods, and analysis topic was covered in two sessions during the workshop. Talks and discussions of both sessions are summarized in this section. The opening talk on the first day was given by **James Sethian,** from Lawrence Berkeley National Laboratory, on the Center for Advanced Mathematics for Energy Research Applications (CAMERA) project. He reminded us that the US Department of Energy (DOE) supports a wide spectrum of experimental science aimed at providing the fundamental advances needed to meet the nation's energy, environmental, and national security challenges. Applied mathematics can play a pivotal role in these investigations. Modest investments have an opportunity to create sophisticated, state-of-the-art mathematics that transforms experimental science and help further discovery.

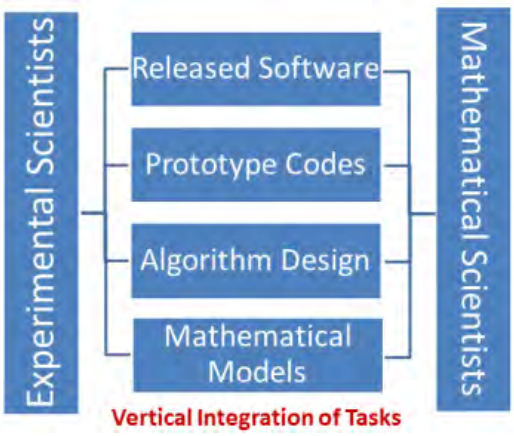

Figure 5. Collaboration is needed between applied mathematicians and scientists on future research.

Fundamental computational methods are needed to extract information from "murky" data, interpret experimental results, and provide on-demand analysis as data are generated. Advanced algorithms can examine candidate materials that are too expensive and time-consuming to manufacture, rapidly find optimal solutions to energy-related challenges, and suggest new experiments for discovery science. New and innovative mathematics can provide tools that will, for example, reconstruct structure and properties from synchrotron experiments, predict behavior of new materials at the nanoscale, direct the hunt for new materials for batteries and gas separation, and optimize steps in the production of biofuels. The required research reaches across traditional boundaries. Building these new enabling technologies requires laying the groundwork through a close collaboration between applied mathematicians and scientists for research aimed at relevant scientific problems that can enhance current and future experiments (Figure 5). Models need to be formulated, equations need to be derived, algorithms need to be proposed, prototypes need to be built, and useable workhorse codes need to be delivered.

CAMERA was formed to meet these needs. Its mission is to develop, weave, and integrate experimental technologies, mathematical algorithms, and advanced computing in tandem. CAMERA has two goals: (a) accelerate the application of new mathematical ideas to experimental science: We are seeing brand-new mathematics that can be directly used to analyze results of experimental research. Traditionally, it takes considerable time for these new ideas to migrate to user communities. By bringing mathematicians and experimentalists together, CAMERA accelerates the early adoption of new mathematics. (b) Provide a broader view: Existing computational techniques are often tailored to specific needs. Approaches may have reached their limit and cannot easily be extended to complex problems with different requirements. CAMERA aims to widen perspective and devise new, more general models and algorithms.

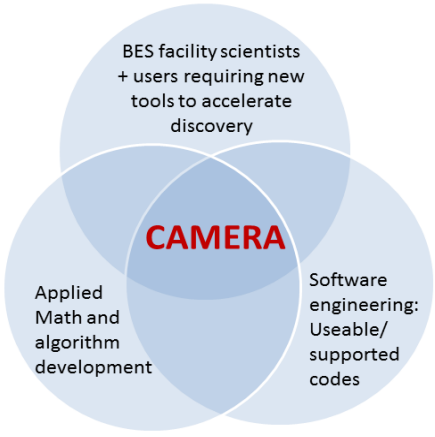

Figure 6. The multidisciplinary approach of CAMERA.

How CAMERA works: CAMERA assembles teams of applied mathematicians, experimental scientists, computational physicists, computer scientists, and software engineers (Figure 6). The teams focus on a particular application area, and participants are typically part of multiple teams. These teams act at the intersection of mathematical and

algorithmic research, focused needs of experimental facilities, and production of high-performance "best-practices" software that is useable by the external community and supported. Building effective teams has challenges and requires considerable work to overcome language/culture barriers between experimentalists and mathematical scientists. An essential element of CAMERA's success is the locality of these teams, which fosters a rapid exchange of ideas that leads to new advancements. Being able to "matrix in" team members for substantial support during the lifetime of a project is crucial to accelerating innovation and leads to understanding across multiple fields.

Current Topics: CAMERA's list of topic areas is constantly growing. Currently, they include work in such areas as ptychography, grazing incidence small-angle scattering, fluctuation scattering, single-particle imaging, electron density and DFT methods, chemical informatics , and tomographic reconstruction. The range of mathematical and algorithmic techniques employed to tackle these problems is substantial and includes computational harmonic analysis; PDE-based techniques for image segmentation; graph theoretic approaches; dimensional reduction and manifold embedding; diffusion maps and nonlinear tensor schemes; sparse and compressed approximation methods; and advances in operator decompositions, machine learning, and statistical methods.

For each of these topics, CAMERA's approach is to perform the fundamental mathematics research, design new algorithms, develop prototype codes, and release software to meet needs of DOE experimental facilities. It executes full vertical integration, so that state-of-of-the-art mathematics is quickly transformed into useable software tools.

Software currently released by CAMERA:

- **SHARP-CAMERA**: SHARP-CAMERA is a versatile package for ptychography, a powerful imaging technique that combines diffraction and microscopy. SHARP (Scalable Heterogeneous Adaptive Robust Ptychography) is designed around advanced acceleration algorithms for convergence and analysis.
- **HipGISAXS:** HipGISAXS is an extensible, high-performance code to execute grazing-incidence small-angle x-ray scattering (GISAXS) analysis. HipGISAXS is used regularly to compute scattering effects from structures such as dense storage media, electrochromic windows, battery electrolytes, OPV BHJ materials, and small molecules assembly.
- **Zeo++**: Zeo++ analyzes and assembles crystalline porous materials. It performs geometry-based analysis of structure and topology of the void space inside a material, alternates or assembles structures, and generates structure representations for use in structure similarity calculations.
- **PEXSI:** The Pole Expansion and Selected Inversion (PEXSI) method is a fast method for electronic structure calculation based on Kohn-Sham density functional theory. It efficiently evaluates certain selected elements of matrix functions (e.g., the Fermi-Dirac function of the KS Hamiltonian), which yields a density matrix. It can be used as an alternative to diagonalization methods for obtaining the density, energy, and forces. It can regularly handle systems with 10,000 to 100,000 electrons.

- **QuantCT and F3D**: QuantCT and F3D are ImageJ/Fiji plugins for image enhancement, filtering, segmentation, and feature extraction from samples imaged using microtomography.
- **MTIP**: Multi-Tiered Iterative Phasing (MTIP) is a new mathematical and algorithmic technique to solve reconstruction problems associated with fluctuation correlation scattering and single particle imaging.

There are several reasons why these combined instrument-related reduction and analysis problems are interesting to mathematicians and to funding agancies to fund. First, for many of the specific science cases, the main takeaway from all of these projects is that knowing what to build, how to build it, and how to use it requires more than a single individual. Thus it is important that cross-disciplinary teams work on each problem. During the discussion it came out that problems being worked in CAMERA are of sufficient interest to the Mathematical community that many people are willing to contribute to solutions purely for the satisfaction of helping to solve the problem.

Travis Humble, from Oak Ridge National Laboratory, gave a short talk and pointed out that everyone has a big data problem—the intrinsic features of the data make each problem different. This includes context, acquisition, and management as well as intent, distribution, and integrity. For experimental physical sciences, data are most often used to validate theoretical models and to inform future choices (e.g., developing applied technologies or planning new experiments). Data are often shared over a wide range of collaborators and must be tagged and tracked to ensure authenticity. Within this context, several new methods for data processing are available for integration with future high-throughput experimental user facilities. They include data processing at the edge of the network, compressive processing methods, and dimensionality reduction to improve acquisition. New methods in artificial intelligence, machine learning, and pattern recognition as well as automated model-based testing can offer robust approaches to post processing big data sets. Validating theoretical models against experimentally compiled big data sets will require equally larger compute systems to perform numerical simulations. Future HPC systems, including those based on novel platforms like quantum computing, can offer substantial jumps in capability over conventional trends in HPC power. The ultimate capability for any system to handle big data will depend on these design choices with natural tradeoffs in size and precision being made.

Rick Archibald, from Oak Ridge National Laboratory, focused his talk on mathematics developed to help with the challenges faced by the DOE experimental facilities. He discussed sparse sampling methods and fast optimization developed specifically for neutron tomography and optimization of neutron-scattering experiments. Sparse sampling has the ability to provide accurate reconstructions of data and images when only partial information is available from measurement. Sparse sampling methods have demonstrated to be robust to measurement error, and we have developed fast algorithms with increased error tolerance for experimentally measured neutron data. These methods have demonstrated the ability to scale to large computational machines on large volumes of data. The methods were developed under the project ACUMEN (Accurate Quantified Mathematical Methods for Neutron and Experimental Science, a project supported under the Applied Mathematics program at the DOE that is focused on developing mathematics for the challenges face by the DOE experimental

facilities. The talk was followed by a lively discussion around the databases. There is a concern that every facility will have its own database and that they will not talk to each other. Some proposed using existing databases, but there is concern that they are too simple for the problem at hand. A proposal was to engage data scientists who are familiar with unstructured data.

The next talk was given by **Ian Foster,** from Argonne National Laboratory. He showed three examples of linking the Advanced Photon Source with the Argonne Leadership Computing Facility to solve the inverse problem. They were single crystal diffuse scattering, x-ray nano/microtomography, and near field high-energy x-ray diffraction microscopy. In the last case they were able to catch errors during a run that typically would not have been noticed until the user returned home. He also emphasized that the data must be accessible in a seamless way to the computer that needs the data. He thinks the solution is long-term (5-year) funded collaborative projects. He also thinks that the computing needs are becoming untenable for the facilities to support.

The final talk in the first part of the session was given by **Jack Wells,** from Oak Ridge National Laboratory, entitled "Integrated Compute and Data Science at DOE's Leadership Computing Facility." After reviewing high impact science enabled by the OLCF, Jack made comments on the integration of compute and data requirements. Specifically, the HPC facilities are being upgraded and there historically has been great synergy between application readiness and early science. Going forward, much more attention should be paid to the portability between architectures. He is concerned that the neutron and x-ray early science agendas for exascale computing are not articulated. The discussion after this talk showed a difference in opinion from the compute scientists and the experimental scientists. On one side the idea was to allocate the experimental beam time according to the computing resource schedule. The other is to allocate the computing resource according to the beam time schedule.

The second part of this session was opened by a plenary talk by **Simon Billinge,** from Columbia University, on the complex materials structure problem. A critical issue in modern high-performance materials, both under development and in production, is that they tend to be complex and heterogeneous, exhibiting important structures on the atomic, nano and mesoscales, up to the macro-scale. These materials challenge our ability to build accurate models to describe their structure and behavior, but worse than that, models that we build are rarely, if ever, validated against data, and even then validation is generally done in a cursory way, by testing against a small number of data points (for example, bulk modulus). This is a show stopper if we want to design materials with tailored properties, a dream of Materials Genomics. The solution to this problem has applied mathematics at its heart, but new approaches that scale to the size of the dimensionality of the problem and that incorporate information from underlying physical models are missing. The complex materials structure problem has at its heart information theory: models of complex multiscale structures have extremely high dimensionality, on the order of three times the number of unique atoms. As structural complexity increases, this rapidly approaches order 10^6 or higher. At the same time, the information content of the data that we have available, for example, from scattering experiments, goes down. This quickly results in materials inverse problems that are ill-posed: the information constraining the solution is less than the degrees of freedom of the model. Approaches are needed to regularize this nanostructure inverse

problem, allowing for robust structure determination and materials design. In this case, we may combine, or complex, heterogeneous information sources coming from complementary datasets, but also with constraints coming from underlying physics models and materials performance criteria specified as inputs, something we call “Complex Modeling and Optimization” or “multi-modal modeling” that is illustrated in Figure 7.

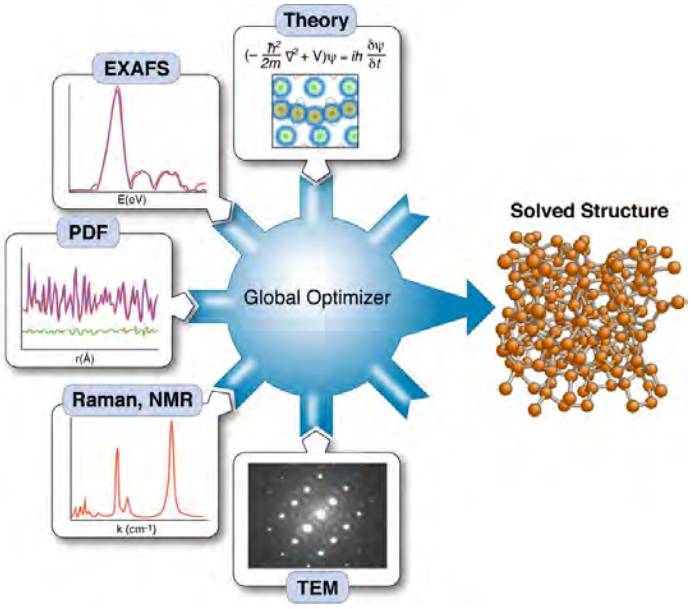

Figure 7. Schematic of the Complex Modeling Optimization paradigm of combining heterogeneous data sources to constrain a unique solution to marginally posed inverse problems, in this case the nanostructure inverse problem.

This is a high-dimensional multiphysics inverse problem under uncertainty, since different (necessarily approximate) physical models are needed to describe each dataset or physical property of interest and to describe the physical model on different spatial and length-scales, and it is not clear how to propagate uncertainties through the different multiphysics models. There are the challenges of correctly handling heterogeneous data sources under uncertainty to obtain the desired result and subsequently assessing our level of confidence in the resulting models (for example, their uniqueness). There are also challenges of how these uncertainties affect the relation between property and structure, how to quantify the uncertainties when encoding information from the different information sources and how to automatically integrate models with different levels of complexity under uncertainties for materials that are themselves heterogeneous and multiscale.

The current situation is somewhat shambolic, and gains that can be made in this area can be expected to have a large impact. The US government (principally DOE) has made significant investments in increasingly powerful measurement tools, primarily for the latest synchrotron based x-ray and neutron sources and ultrahigh-resolution electron sources, all of which provide unprecedented quantities and qualities of scattering and spectroscopic information. However, it is safe to say that most of the data is thrown away, and certainly not used to its full potential. Mostly it is not stored in a machine-readable format or coupled with metadata suited to data mining. There are very few efforts to combine data from different experiments in a concerted way to quantitatively constrain solutions to materials inverse problems. In large part, this is because we do not know how best to encode the information content in the data, nor have we explored how to combine the information effectively to constrain models.

The first short talk was given by **Greg Schenter,** from Pacific Northwest National Laboratory, who talked about the need to enhance our understanding of fundamental molecular phenomena. The goal is to establish a connection between molecular simulation and observable signals from measurement. The approach that we use consists of defining a connection between a molecular system, a description of molecular interaction, and statistical mechanical simulation. These techniques are used to generate an appropriate ensemble coupled with the calculation of an observable signal from molecular simulation, which in turn, is compared with an observable signal from measurement. This approach becomes more significant as systems and phenomena become more complex. Complexities that we consider consist of inhomogeneous systems, interfaces, and chemically reactive systems. Interfaces between phases (gas/liquid,liquid/liquid, liquid/solid, and gas/sold) are prevalent. There is a continuum of sophistication that connects theory and experiment. It is necessary to establish a balance. Decisions must be made, and decisions made in despair must be avoided. This requires an honest assessment of capabilities. Frameworks (methods) must be "used properly" to have predictive power, keeping in mind that one size does not fit all. It is effective to build a hierarchy of frameworks that are self-consistent. We need O(N), $O(N^2)$ and $O(N^3)$ codes that are integrated, representing various levels of accuracy.

The next talk was given by **Ray Osborn** from Argonne National Laboratory. As part of an internally funded project, Discovery Engines for Big Data, they have been developing flexible methods of handling large data volumes that might be generated at multiple facilities. In particular, they have developed computational tools with the goal of enabling a joint analysis of single crystal diffuse x-ray and neutron scattering, using a prototype framework that can be applied to other measurement techniques at the APS and other large scale x-ray and neutron facilities. These tools are designed to be used by both instrument scientists and facility users, allowing them to collect, visualize, and analyze "big data" without requiring specialized expertise, other than some basic knowledge of Python. In experiments so far at the APS and CHESS, raw images from fast area detectors have been streamed to a remote server and automatically stacked in NeXus files by Python scripts, which also harvested the relevant instrumental and sample metadata. By registering these files in the newly developed Globus Catalog, the data were immediately available for remote visualization and analysis, using an extensible Python GUI, NeXpy (http://nexpy.github.io/nexpy). This loads NeXus file trees so that all of the data and metadata are accessible at a granular level using Python Remote Objects. It is then possible to write simple scripts to process the results in real time during the experiment from any location without needing a fast network, even with data sets of several hundred GBs. By using high-level "wrapper files," which contain pointers to large data sets, which could, in the future, be identified by a global URI, data from multiple facilities, and even theoretical simulations, can be encapsulated in a single portable file. Scientists from the same experimental team, who may be performing different modes of analysis on the same data, e g., powder diffraction or PDF analysis, can copy these wrapper files, which are typically only a few MB in size, from the data servers and customize them for their particular application, with each able to access the raw or processed data using the Python Remote Object protocol. These tools will shortly be tested at the SNS, where it will be used to access data reduced by the Mantid framework (http://www.manitdproject.org), and it is believed that experience gained on this project should be of use in the design of remote data facilities.

The final short talk, by **John Tranquada,** from Brookhaven National Laboratory, focused on dynamic correlations in the absence of order. He stressed the value of neutron-scattering data and raised questions related to the integration of theory and experiment: How could data obtained from neutron scattering be shared with theorists? Is there a general way to parameterize the results? Is there a better way to help the experimental and theory teams discuss these results?

DISCUSSION

A number or participants pointed out that in other fields (e.g., recent astronomy experiments), software represents a significant part of the total project cost. Similarly, in industry, a large part of the development cost for instrumentation such as a magnetic resonance imaging machine is spent on developing the control and data analysis software. Advances in instrumentation at scattering user facilities as well as available computation power require adopting a similar approach in the study of materials. Making this transition requires funding and, most importantly, the formation of interdisciplinary teams composed of domain scientists, computational scientists, applied mathematicians, theorists, and data scientists. On the hardware side it was agreed that the community needs a complete ecosystem, from institutional clusters to current and future leadership computing facilities. Looking into the future, the promise of a petascale level computer in a rack makes is feasible to have the computational power of today's leadership computers in a university department or at beamlines at the neutron and light sources. The common themes coming out of this session are as follows.

- Diverse teams of instrument scientists, mathematicians, software specialist, and hardware specialists are needed to solve the current complex problems.
- There is a need for a hierarchy of solutions to most problems. Some faster and/or focused, some all-encompassing, but slower, and everything in between.
- Compute environments are becoming ever-more heterogeneous one should be ready for them.
- Easy access to the data by the appropriate compute platform is essential.
- Useful access of the data to teams that did not take the data (e.g., theorists) is becoming a need.

APPENDIX I: WORKSHOP AGENDA

OAK RIDGE National Laboratory

AGENDA

MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

Frontiers in Data, Modeling, and Simulation

Chair, Peter Littlewood

Argonne National Laboratory

March 30–31, 2015

Time	Event
Monday March 30, 1015	
8:00–8:30 a.m.	Breakfast
8:30–9:00 a.m.	**Welcome and Charge** Peter Littlewood (ANL), Thomas Proffen (ORNL)
9:00–10:30 a.m.	**Hard and Quantum Materials** **Keynote speaker: Tom Devereaux, SLAC** Chair: Colin Broholm, Johns Hopkins **Short talks**: Maria Fernandez-Serra, Gabi Kolliar, Thomas Maier
10:30–11:00 a.m.	**BREAK**
11:00 a.m.–12:30 p.m.	**Data, Math, Methods, Analysis and Computing I** **Keynote speaker: James Sethian, UC Berkeley** Chair: Ian Foster, ANL **Short talks:** Travis Humble, Rick Archibald, Ian Foster, Jack Wells
12:30–1:30 p.m.	Group photo / Discussions / Lunch **Chairs: Thomas Proffen, Peter Littlewood**
1:30–3:00 p.m.	**Materials by Design** **Keynote speaker: Thomas Schulthess, ETH Zurich** Chair: Bobby Sumpter, ORNL **Short talks:** Timmy Ramirez-Cuesta, Sasha Balatsky, Mike Norman
3:00–3:30 p.m.	**BREAK**
3:30–5:00 p.m.	**Soft Materials** **Keynote speaker: Monica Olvera de la Cruz, Northwestern** Chair: Matt Tirrell, U Chicago **Short talks:** Bobby Sumpter, Matt Tirrell, Fyl Pincus
6:00–9:00 p.m.	**Dinner / Location: Argonne Guest House** **Neutron Sciences: Present and Future** **Speaker: *Alan Tennant - STS, other workshops…***

OAK RIDGE National Laboratory

AGENDA

MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

Tuesday, March 31, 2015	
8:00–8:30 a.m.	**Breakfast**
8:30–9:00 a.m.	**High Performance Computing, a Neutron Scatterers Perspective** **Speaker: Colin Broholm, Johns Hopkins**
9:00–10:30 a.m.	**Data, Math, Methods, Analysis and Computing II** **Keynote speaker: Simon Billinge, Columbia U** Chair: Garrett Granroth, ORNL **Short talks:** Greg Schenter, Ray Osborn, John Tranquada
10:30–11:00 a.m.	**BREAK**
11:00 a.m.–12:30 p.m.	**Bio Materials** **Keynote speaker: Benoit Roux, U Chicago** Chair: Paul Langan, ORNL **Short talks:** Loukas Petridis, Pratul Agarwal
11:30 a.m.–12:30 p.m.	**Panel:** User Outreach
12:30–2:30 p.m.	**Discussions / Lunch / Preliminary report writing / Breakout as needed** Chairs: Thomas Proffen, Peter Littlewood
2:30–3:00 p.m.	**Wrap-up**

APPENDIX II: LIST OF PARTICIPANTS

Name	Institution
Agarwal, Pratul	Oak Ridge National Laboratory
Archibald, Rick	Oak Ridge National Laboratory
Balatsky, Sasha	Los Alamos National Laboratory
Billinge, Simon	Columbia University
Broholm, Collin	John Hopkins University
Devereaux, Tom	Stanford Linear Accelerator Center
Fernandez-Serra, Maria Victoria	Stonybrook University
Foster, Ian	Argonne National Laboratory
Granroth, Garrett	Oak Ridge National Laboratory
Humble, Travis	Oak Ridge National Laboratory
Kotliar, Gabi	Rutgers University
Littlewood, Peter	Argonne National Laboratory
Maier, Thomas	Oak Ridge National Laboratory
Norman, Mike	Argonne National Laboratory
Olvera de la Cruz, Monica	Northwestern University
Osborn, Ray	Argonne National Laboratory
Petridis, Loukas	Oak Ridge National Laboratory
Pincus, Fyl	University of California Santa Barbara
Proffen, Thomas	Oak Ridge National Laboratory
Ramirez Cuesta, Timmy	Oak Ridge National Laboratory
Roux, Benoit	University of Chicago
Schenter, Gregory	Pacific Northwestern National Laboratory
Schulthess, Thomas	ETH Zurich
Sethian, James	Lawrence Berkeley National Laboratory
Sumpter, Bobby	Oak Ridge National Laboratory
Tennant, Alan	Oak Ridge National Laboratory
Tirrell, Matt	University of Chicago
Tranquada, John	Brookhaven National Laboratory
Wells, Jack	Oak Ridge National Laboratory

APPENDIX III: INVITATION LETTER

Dear Colleague,

As part of the thought process to identify the needs of the scientific community in the areas of Neutron Science and possible areas of cooperation with Photon Science, we have been undertaking workshops to identify the Science Grand Challenges for the next decade. Workshops have been held in four complementary topics: Quantum Condensed Matter (at Lawrence Berkeley National Laboratory), Biological Systems (at University of California, San Diego), Soft Matter (at University of California, Santa Barbara), and Materials Discovery (in Chicago). These workshops have successfully determined where neutron and other experimental scattering probes complement each other and the most compelling science challenges for the next decade and beyond. However, a key finding of this activity has been the recognition of the central role that high performance computing and big data will play in both experiment design, but also in data modeling and simulation. As continuation of this DOE process to help in defining the future course of these user facilities, we are holding a workshop dedicated to this topic entitled "Grand Challenges for Neutrons and Supercomputing". In order to facilitate deeper interactions, this workshop is limited to about 40 participants and is by invitation only.

With this letter, we are inviting you to join us in defining the future needs in data, modeling, and simulation as relevant to all aspects of neutron sciences. We are planning to hold the workshop March 30–31, 2015, at Argonne National Laboratory. There will be no registration fee, and local arrangements will be covered by the workshop. Travel will be funded and arranged by the Oak Ridge National Laboratory. In order to facilitate the logistics of organizing the workshop, please send your response to this invitation to Toni Sawyer (sawyertk@ornl.gov). We would appreciate receiving your response as soon as possible, but no later than Friday, March 6, 2015.

We look forward to a vigorous and thought-provoking workshop.

Best wishes,

Peter Littlewood

Workshop Convener

Thomas Proffen

Workshop Facilitator

APPENDIX IV: ACKNOWLEDGMENTS

The workshop organizers would like to thank *Toni Sawyer* (ORNL) and *Lupe Franchini* (ANL) for taking care of the entire workshop logistics and making this workshop a great success.