



Report of the MPS AC Subcommittee, MPS and the Living World (MPSLW)

**Transforming the Future of Biotechnology
January 2022**

TABLE OF CONTENTS

I. SUMMARY	3
II. INTRODUCTION	4
III. EXPERIMENTAL TOOLS	6
IV. EMERGING THEORIES	9
V. CRITICAL APPLICATIONS AND PROBLEMS IN BIOTECHNOLOGY	16
VI. HUMAN-BIOTECHNOLOGY INTERFACE	18
VII. MPS AND THE LIVING WORLD SUBCOMMITTEE	20
VIII. APPENDIX - WORKSHOP DETAILS	21

I. Summary

A transformation in biological research requires “audacious” goals.

The NSF fosters transformational research by supporting fundamental science and engineering needed to accelerate breakthroughs in the development of powerful tools and revolutionary advances in chemistry, physics, mathematics, and computer science domains that underpin the living world. Increasing NSF’s investment over the next two decades at the interface between MPS and the living world will have a dramatic impact on the future of biotechnology.

The Subcommittee met from January to August in 2021, held discussions with diverse experts in the field, conducted surveys of the research community, and organized a virtual Future of Biotechnology Workshop in June 2021 (see Appendix for Workshop Agenda). This report summarizes the findings, deliberations, and recommendations of the Subcommittee resulting from these activities.

II. Introduction

Biotechnology is a broad discipline in which biological processes, organisms, cells, or cellular components are harnessed to develop new technologies. New tools, techniques, and products developed by biotechnologists are increasingly employed to accelerate progress in industries from agriculture to medicine, from electronics to energy, from advanced manufacturing to digital communication, and many others.

In October of 2020, MPS and the Living World was constituted as a Subcommittee of the Mathematical and Physical Sciences Advisory Committee (MPS AC) to deliberate and to provide recommendations on how best to:

- Harness all the capabilities of MPS to enable and transform the future of biotechnology (biotech)
- Build transformational collaborations across MPS and beyond
- Advance research in industries from Agriculture to Medicine, Electronics to Energy, Advanced Manufacturing to Digital Communications, and Biotech.

Specifically, the Subcommittee was charged to explore the opportunity space and recommend both short- and long-term strategies for MPS that provide insights into the following questions:

- What are the fundamental science questions that, if answered, could significantly accelerate future biotechnologies?
- Which of these science questions are unique to the MPS communities? And what new tools and/or techniques are needed to address these science questions?
- How could convergence among MPS disciplines advance the field of biotechnology?
- How can partnerships with other parts of NSF, with other agencies, and with private foundations and/or companies advance these goals?
- In addition to MPS contributions to the development of fundamental knowledge in biotechnologies, what areas of biotechnology might reciprocally advance fundamental science in AST, CHE, DMR, DMS, and PHY?

The Subcommittee organized itself into subgroups for its work, identifying overarching themes, existing gaps, and best practices in four focus areas:

Four Focus Areas:

1. Experimental Tools (*Nebojsa Duric, Ka Yee Lee, Jeff Pixton, Dagmar Ringe, Eranthie Weerapana*)

Breakthroughs in most scientific fields are often enabled by new tools for the design, fabrication, and analysis processes. Identification of existing gaps in the toolset for biotechnology, spanning a wide spectrum from synthetic to analytics to computational ones, and facilitation of such tool development, can play an important role in transforming the future of biotechnology.

2. Emerging Theories (*Rommie Amaro, Hebert Levine, Michael Murrell, Neal Woodbury*)

A scientific theory provides a framework for comprehending an aspect of the natural world, and emerging theories stand to open up new avenues for our understanding of observed phenomena, reorder old knowledge into a new framework, and render progress possible in solving problems stymied by the earlier framework. Identifying existing gaps in theories in the chemical, physical, or biological areas, and supporting their closure can have a transformative effect on the advancement of the field.

3. Critical Applications and Problems in Biotechnology (*Sara Del Valle, Moh El-Naggar, Trachette Jackson, Jennifer Lewis*)

New tool, techniques *and theories* developed in biotechnology are increasingly employed to make progress in industries from agriculture to advanced manufacturing to medicine. What are the critical applications and problems that stand to benefit from future biotechnologies? What are the outstanding issues in these areas, and what kind of biotechnological innovations are needed to help advance these critical issues?

4. Human-Biotechnology Interface (*Virginia Cornish, Chaitan Khosla, Sherine Obare, William Tolman*)

Technological innovations no doubt drive progress. However, with an increasing number of these innovations sitting at the human-technology interface, new technologies can also raise ethical questions, and have the potential of posing new risks to society and environment. Moving forward, responsible biotechnological innovations would require an interdisciplinary approach involving other disciplines such as ethics and social sciences.

III. Experimental Tools

Biotechnology requires a multiscale approach, as important systems and issues range from molecular level on up to ecological milieu. Perhaps the most critical area of needed progress is at the cellular scale, arguably where physics and chemistry first combine to create processes that are truly biological. The cell functions with complicated and interrelated pathways. In model organisms, many of these pathways have been described by the biology community, especially with regard to their constituent components. However, how these pathways relate to each other is more difficult and essential to an understanding of the workings of the cell as a whole, to an understanding of the workings of the cell in response to internal or external stimuli, and to the workings of an organism composed of cells. Approaches that would enable more quantitative descriptions of functional behavior and would therefore allow for the prediction of disruptions is essential for a wide range of biotechnological purposes, such as individualized medicine and potential genetic intervention, and environmental desiderata such as bioremediation.

To achieve these goals, the mechanisms that govern how cells work must be understood. Quantitative models of the inner workings of a cell are needed, which can only be achieved by integration of theory and experiment. We first address imaging to illustrate areas where the MPS community can participate in making essential progress. Important experimental techniques in biotechnology include cryo-EM, advances in proteomics especially at the single-cell level, metabolomics, accurately measured kinetic parameters, and fine-scale resolved imaging. New computational methods are needed to combine the results from such studies into a set of coherent models, as discussed below.

3.1 Tools for Cell and Tissue Imaging: Microscopy is a powerful tool that is critical to furthering our understanding of the function and organization of metabolites, biomolecules, and organelles within the cell, as well as the complex interplay and communication that occurs between multiple cell types and the surrounding microenvironment at tissue and organismal levels.

Imaging tools can be improved through:

- *Advances in small-molecule fluorophores:* improved photostability, switchable properties for super-resolution imaging, and linkage chemistry to incorporate fluorophores into biomolecules
- *Improved protein fluorophores:* rational engineering and library screening to improve photophysical properties, maturing time, and in-vivo stability

- *Functional probes*: improved biosensors that monitor cellular processes such as metal influx/efflux, electrochemical gradients, redox potentials, etc.
- *Increased spatial and temporal resolution*: <10 nm spatial resolution (requires advances in both fluorophore design and instrumentation)
- *Multiplexing*: analyzing 10-100 biomolecules concurrently (e.g., using DNA-barcoded antibodies)
- *Correspondence between DNA/RNA sequencing and imaging*: better integrate available single-cell sequencing tools with single-cell imaging
- *Deep tissue 3D visualization*: requires advances in instrumentation (e.g., light-sheet microscopy) and sample preparation (e.g., sample clearing and hydrogels)
- *Advances in data analysis*: necessary for entire imaging pipeline, and can be integrated with machine-learning to obtain the maximal information with the minimal amount of imaging

Potential applications of these improved imaging tools include:

- *Neurobiology*: spatial mapping of neurons to better understand long-range communication in the brain
- *Pathology*: imaging of 3D tissue biopsies
- *Nuclear organization and epigenetics*
- *Cellular signaling*

3.2 Other areas: There are many other areas where MPS-enabled progress in biological tools will become increasingly critical. The DNA genome, RNA transcriptome, and microbial genomes (DNA or RNA) contain numerous, but rare chemical modifications that are critical to organism function. Such modifications punctuate the genomes and choreograph everything that happens in the cell. Therefore, developing techniques that permit sequencing DNA and RNA that can detect these modifications is essential, allowing one to understand how small but crucial genomic and transcriptomic changes direct biochemistry – every molecular action – within a given cell.

Significant gaps in our understanding must be bridged to enable targeted design of artificial enzymes that catalyze unnatural reactions, especially the understanding of what gives amino acids specific catalytic power in the context of a protein. In all cases, there is a need for the application of principles from physics, especially, to simulate the relationships between physical phenomena which lead to control mechanisms, communication, and catalysis.

Nondestructive temporal and spatial measurement analysis *cycles* are critical to continuous monitoring in medical, environmental, and agricultural fields. Many -omics approaches require the destruction of the biological sample (e.g., mass spectrometry) and thus the sample can be analyzed at only a single time point, and without spatial information. A key element to the continual development of omics is to therefore to quantify non-destructively, and measure spatial and temporal information, e.g., high-throughput optical data where possible.

Factors to consider when developing new tools:

- Accelerate the dissemination of tools to the broader community
- Encourage rapid progression to commercialization; applicable to small-molecule probes and instrumentation (entire imaging systems, and components such as lenses)
- Open-source computational tools
- Interdisciplinary collaboration; tool developers (chemists, engineers, computational scientists, statisticians) working together with those with domain knowledge (biologists)
- Training – encourage diverse training of students in chemical, physical, computational and biological sciences
- Encourage communication between tool developers, data analyzers, and biologists from very early on in the development pipeline to better streamline data collection, data analysis, and hypothesis testing.

IV. Emerging Theories

New theories are needed to support emerging experiments aimed at describing living systems and predicting what makes these systems adaptable and able to evolve. Here, the lofty goal is to describe systems at all levels, from the molecular to the organismal and environmental, and therefore predict how these systems function and react to different stimuli as well as how these systems can be controlled and programmed. Close collaboration between experimental design and theory is required to achieve our vision of “biotechnology by design”.

The growing importance of analysis, computation, and theory has been one of the major impacts of the increased efforts of physical scientists to tackle questions arising in living systems. It is important to stress that theory comes in many forms. As biological data continue to grow in quantity and complexity, more powerful analytical methods (advanced “bioinformatics”) will be required to untangle complex experimental observations. This data-driven approach will not only facilitate data interpretation but augmented by machine learning, will also venture into data-driven predictive modeling. This path forward brings its own set of challenges, to be discussed in detail below.

Importantly, this is not the only way theory will impact biology. The history of physical science reveals that the formulation of simplified models focusing on specific, often surprising, experimental facts can help reveal key principles underlying the operation of complex systems. An example of this is the way in which relatively simple polymer models addressing the Levinthal paradox in protein folding, (namely the fact that for an arbitrary polypeptide, there was no obvious way for a unique ground state conformation to be found in physiologically reasonable times), led to ideas of minimal frustration and folding landscapes. Ideas emerging from models regarding optimality in non-equilibrium systems or the positive effects of stochastic fluctuations are less well established but may prove crucial as well. Often simplified models lead to suggestions for key experiments, and often the theoretical insights and constructs built by the analysis of such models informs more quantitatively accurate frameworks, whether hypothesis or data driven. These mathematical models can be thought of as the much more powerful versions of the cartoons that grace most biology papers and textbooks. Adapting a common aphorism, simplified models are always wrong, but some simplified models are essential. Finally, it is worth noting that the

term "simple" can be misleading. Stochastic nonlinear dynamical models of tens of interacting genes or coarse-grained representations of the polymer interactions underlying remodeling of the cellular cytoskeleton are simple as compared to the true depths of biological complexity, but yet require major computational efforts and advanced statistical methods to determine their behavior.

4.1 Mining of multidimensional data sources will be increasingly important in all of the areas described above. The goal of mining is to identify motifs within the data that are predictive of cellular or physiological outcomes. Contemporary and commonly used approaches include methods to reduce this parameter space to the degrees of freedom that most accurately represent the data (e.g., decomposition) or in identifying motifs within the data through pattern recognition (e.g., feature learning). There are limitations in data acquisition as well as in algorithmic approaches. Dynamic and spatial data and analysis will be key. Data mining approaches will need to consider the cost of computational complexity – algorithms need to quickly, and efficiently identify the most important degrees of freedom. Theoretical approaches will need to be developed to both efficiently mine existing data, but also apply physics-based and information-theoretic approaches to constrain large or incomplete data sets. Overall, the appropriate application of theory should enable better predictive power with less data. This may include combining developed models in control theory (e.g., dynamical systems) where incomplete information is ‘coarse-grained’ in identifying an input / output relationship, with new physics-based theories (e.g., non-equilibrium statistical physics) to apply energetic bounds that constrain system interactions.

4.2 Control theory. Traditionally used in engineered systems, control theory identifies an objective – an output robust to perturbation of upstream components, modulation of a delay between input and output signals, and general stability analysis. With this type of approach, interacting systems of biochemical molecules form modules, whose interactions are defined by a transfer function – the relationship between inputs and outputs. An enabling feature of control theory is its ability to deal with non-linear systems, ubiquitous in biology. As unknown parts of a system can be ‘coarse-grained’ and treated as a black box with its own inputs and outputs, control theory may be less constrained by needing all observable degrees of freedom in a system, thus making it predictive with less information. Examples include control in the cardiovascular system, in which external workload (exercise) is related to heart rate, representing the non-linear relationship between homeostasis and metabolic efficiency. Relatedly control models have been used to demonstrate the tradeoffs between robustness and efficiency during glycolytic oscillations.

4.3 Efficient data collection and analysis cycles will be critical because the parametric space outlined by omics approaches may generate more data than actually needed to make useful predictions. Thus, the potential to extract information efficiently may be more useful than acquiring more data, highlighting the need for more theory rather than more technology in certain applications. Unfortunately, contemporary models of machine learning require too much stimuli to be predictive of complex phenotypes. The human brain can, for example, use as little as a single image as an input to subsequently recognize similar patterns. Thus, there is a need to optimize machine learning approaches with other principles that may enable predictability with less test data. One such approach is to use theory to combine machine learning approaches with energetic constraints or 'tradeoffs'. For example, in learning to recognize dynamical motifs in stochastic time-resolved data, the concept of thermodynamic irreversibility can be applied to constrain further motif recognition. Namely, that non-equilibrium systems (e.g., systems of biochemical reactions) are defined by an imbalance in forward and reverse processes. Again, if time-based data can be acquired, then the irreversibility in time may be used to constrain the possible dynamics in the system, thereby reducing the total number of possible interactions dramatically. Alternatively, there may be principles by which other measurable properties of the system, including efficiency or rate of a process exists in balance (trade off) with irreversibility, and thus may further constrain system dynamics.

4.4 Merger of first principles theory with data intensive analytical methods is likely to be required. One needs an overarching framework, developed from first principles, which connects to omics-based parameterization. Example theoretical work on this topic includes defining a non-equilibrium dissipation of energy and suggesting that non-equilibrium systems choose biochemical parameters in order to optimize or extremize this value. Alternatively, one might imagine directly converting concentrations and activities into physical parameters such as forces and flows, thus converting the language of biochemical rates and activities into random variables and fields. With fields for example, one applies principles of mechanical equilibrium in cases where principles of energy conservation or variation cannot be found. In either case, as we assume that the complex interactions of cellular components are constrained by the laws of physics, we should apply these constraints to enable enhanced predictability. Continued effort in measurement and data acquisition alone may never provide this level of predictability.

4.5 Development and validation of emerging theories will be based on a variety of data streams. Medical data will be increasingly personal, continuous, and pervasive. In-body and on-body sensors, smart environments, IoT tracking, consumption tracking, location tracking (exposure), external and internal imaging, voice analysis, motion analysis will all likely play an

increasing role. Since epigenetics constantly modifies gene expression and DNA sequences themselves are not static and homogenous throughout all cell types (B-cells and T-cells, for example), one must collect these data as well. Immune profiles (the billion different immune cells monitoring and constantly evolving to meet new threats), transcription profiles and proteomics (the current state of gene expression) are huge information sources and highly dynamic. Integrating this disparate data in real time with medical outcomes and fundamental biological theory to optimize therapies, behavior, consumption, and activity will pose a massive challenge in storage, computation, interpretation and rational decision making.

4.6 Solving Inverse Problems in Support of Biological Studies. While the development of specific methods can help solve specific problems, perhaps a broader, more foundational approach is needed to drive major paradigm shifts over a 20-year time scale. Such an approach recognizes commonalities in the required tool sets and proposes a long-term development plan for such commonalities.

One specific idea in this regard is to emphasize the importance of solving inverse problems to help us better understand the functionality of biological systems. An inverse problem approach consists of generating of a “forward model” that simulates the physical and functional properties of the system being studied, followed by an inversion process that interprets the data in light of the forward model so as to yields images of desired properties. A coordinated collaboration between physicists, mathematicians, and computer scientists, for example, would yield foundational change in solving relevant inverse problems which in turn would drive a major transformation in biological studies.

Current limitations of inverse problem solutions can be understood as follows:

- Imaging tools for biological systems are bounded by the need for data acquisition and image reconstruction times to be shorter than the time scale of system changes
 - o Data acquisition rates are limited by cost of electronics
 - o Image reconstruction time is limited by computing costs
- Such limitations prevent the realization of real time remote sensing of biological systems on macro, micro and nano scales
- Once the biological problem at hand is defined, the specifications for the data acquisition rates and image reconstruction times are determined
- Tools are then needed to meet the specifications
- Currently, such tools are not available for the biological problems we wish to address because the specifications are far beyond current capabilities.

To bridge the gap between current capabilities and the specifications imposed by the biological problem we wish to solve, several areas of breakthrough research and development are needed:

- **Physics:** Imaging tools are implementations of practical solutions to inverse problems. It is therefore critical that such inverse problems be defined with sound physics that efficiently define the problem without any more physics content than needed.
- **Mathematics:** Novel methods for simplifying inverse problems are greatly needed. Once established by physics, mathematical tools are needed to simplify the equations that underpin the inverse problem. Such equations should be as simple as needed to solve a specific problem. Formulating inverse problems in ways that require fewer computation steps is also needed.
- **Computational Science:** Once established by physics and formulated by mathematics in the form of equations, novel equation solvers are needed. The sole purpose of such solvers is to minimize the time required to get to a solution. Novel computation methods are needed to maximize computational efficiency.

4.7 Advancing the theory front. We first focus on specific systems that harbor significant potential for future biotechnology progress yet require complementary theoretical investigation. Next, we discuss other general areas where investments in the development of theoretical concepts and methods could be of high payoff.

Some specific areas that could use more focused theoretical efforts include:

- *Better understand complex microbial communities.* Multi-species microbial colonies are of importance for issues as diverse as health consequences of the microbiome, fouling effects due to biofilms, elemental cycles in the environment and waste remediation by synthetic biosystems. Aside from gathering and traditional interpretation of metagenomic data, there seems to be little in the way of quantitative approaches grounded in fundamental theory. Further, there are unexplored model systems which may present opportunities for translating novel functionalities into engineered systems. One such example is 'extremophiles' in which organisms live in extreme environments traditionally hostile to life (e.g., temperature / pH or resource limited, photosynthesizing in low light conditions).

- *Develop theoretical descriptions of synthetic cells.* Synthetic cells are one way of coming to grips with the question of what is the minimal set of ingredients that define a living organism and thus develop the fundamental theoretical underpinning required in the analysis of more complex cellular systems. Also, synthetic cells could have a wide range of applications. There has been some initial work on computational modeling of the metabolism and genetic information processing of bacteria with minimal genomes, but more is needed to connect with bottom-up experiments. This is currently an area dominated by empirical experimental studies, with limited theoretical insight into such basic questions as: What is the minimum number of genes needed? and What processes are absolutely essential to maintain cell viability. The development of minimal systems may provide novel targets for therapeutics, interventions, and bioprocessing. For example, current theoretical approaches can be used to estimate the maximum efficiency of CO₂ and sunlight conversion to biofuel in engineered microorganisms. In this sense, efficiency has multiple aspects, each benefiting from a theoretical approach. On the cellular level, it relates to how the genome is re-engineered to relate metabolic inputs to target outputs. On a larger scale, theory can be used to understand how organisms can maximize their growth and production.
- *Advance the theory of complex molecular machines.* The assembly and function of multi-component biological systems are prohibitively complex, such as the focal adhesion complex, the post-synaptic density, and the machinery governing DNA folding and transcription/replication dynamics. Do we know enough about simpler aspects of molecular biophysics (folding of single domain proteins, protein-protein binding interfaces) to be able to better understand and re-engineer this type of living nanotech? To describe diverse self-assembling systems (machines) in biology requires building upon traditional models for phase transitions, such as Smoluchowski and Becker-Döring. Further theoretical approaches will need to incorporate assembly of numerous components over disparate time and length-scales.
- *Develop predictive methods for pharmacological treatments* – metrics that can be used to determine the decision of which patients to treat, with what drugs, and with what dose. Thus, these models incorporate multiple types of data, using mechanistic interactions between agents and pathways, and integrates them to make predictions about a clinical outcome. This can enable better drug development and inform on clinical trials. Current avenues for theoretical approaches include predictions of the outcomes of multiple drug treatments and higher order interactions.

The second set of recommended focus areas concern general advances that transcend the application to specific questions.

- *Integrating the design of large-scale data collection with analytical approaches.* Simply collecting data is expensive and inefficient. Direct integration with analysis, determining what data and at one level it should be collected to provide the greatest predictive power for the least resource investment will be critical.
- *Integrate AI and machine learning with fundamental theory.* Machine Learning and AI is everywhere, but most of the practice in this area is highly empirical. Future progress could very well depend on the development of a theoretical framework, merging the constraints of fundamental physics and chemistry with the data sources available in biology, allow more accurate and predictive descriptions of complex, nonlinear biological processes with less data.
- *Explore a greater diversity of biological models and data collection modalities.* Most data collection and theoretical work at the cellular levels and above have been motivated by a small subset of animals, probably because of NIH influence. Thus, there is a gap in understanding the broad diversity of species and in particular plants, some of which are of immediate use for improved agriculture.
- *Incorporate environmental context and multiscale connections.* Whether one is considering the health of an individual, the health of agricultural plant and animals or the health of the biosphere at large, multiscale data, computation and theory will be required to allow accurate predictions and suggest interventions. Molecules, cells, organisms, communities, and the physical environment all need to be considered. Biological and physical data and theory to describe the interactions at all scales is key.

V. Critical Applications and Problems in Biotechnology

We have already mentioned several areas where the integrated development of new experimental tools, data analytics, and theory will lead to tremendous advances in biotechnology. To recap some of these, potential outcomes include:

- building synthetic, programmable macrophages to combat disease
- mapping a protein/enzyme to predict the effectiveness of a drug
- designing a protein to manipulate its function so that it could work better than the native one
- cell-free synthetic biology that enables new discoveries, e.g., new catalysts
- self-regulating synthetic organisms (most likely bacteria) to grow coatings and predesigned structures
- development of a platform to easily design microbes for synthesis in synthetic biology.

Here we focus on overarching grand challenges that could guide NSF investment. Each of these would require multiple advances across convergent fields represented by the MPS, Biology, and Engineering Directorates at NSF.

5.1 Promote research focused on living diagnostics and cellular therapeutics. The emergence of synthetic biology is opening new avenues for engineering gene circuits and/or harvesting cellular constituents for use real-time biosensors. Opportunities abound to expand the synthetic biology toolbox through a better understanding of design rules, creation of new circuit components, and use of transiently engineered cells. A broad range of complementary research activities should be supported that focuses on cell-based therapeutics. Through programming human cells to serve as drugs, biotechnology is ushering in a new era of living pharmaceutical-based products. New research – ranging from design and employ artificial antigen-presenting materials as scaffolds to harnessing small molecules to drive stem cell differentiation – are needed to improve the yield and quality of therapeutic cells. Advances across the living diagnostic and therapeutic space will require implementing design-build-test methodologies augmented via deep learning. Hence, Center-scale investments that bring together expertise across multiple NSF Directorates, including MPS, Biology, and Engineering, are needed to achieve this grand challenge.

5.2 Promote research that advances stem cell reprogramming, 3D/4D imaging, and deep learning methods. There is a growing focus in engineering human tissues and organs for drug discovery and screening, disease modeling, and tissue repair, replacement, and regeneration. To achieve this grand challenge, research advances in programming multicellular building blocks from stem cells with the requisite cell types, function, and architecture are needed. New 3D/4D imaging techniques must be developed that enable single cell mapping deep within native and engineered tissues and organs. Such methods will provide instruction sets for the design and assembly of complex tissue architectures. Finally, theory-based methods should be applied to unravel large, complex data sets that emerge these experimental techniques.

5.3 Increase focus on agricultural health, which directly connects to human health. As the climate changes, we will need to change agricultural practices and likely modify crop genetics. Temperature tolerance, low water tolerance, brackish water tolerance, processes for conserving key nutrients (e.g., phosphorous). Higher density production which will likely mean less reliance on animal products will be needed. More and more we will have to control how we “farm” the oceans and waterways. We will likely have to design plants to create the balance of nutrients needed with the least resource allocation. All this will only be possible through constant monitoring of the agricultural environment and the physiology of the plants and animals involved at all levels from the molecular to the field. As with other areas, there are big challenges in design and logistics of collection, storage, computation, integration, and interpretation of data.

5.4 Increase focus on monitoring the biosphere will become ever more critical as we adapt to and attempt to mitigate aspects of climate change. We will need to have significant data at the population level for both humans and other species across the planet, including population numbers, diversity, DNA sequences and changes in genetics, epigenetic profiles, gene expression (RNA, protein), pathogen spread, migration, molecular profiles of different kinds. Physical parameters will need to be measured: temperature, rainfall, fires, water purity, atmospheric conditions, dynamic details of the molecular make up of water, air, soil across the globe. Satellites, drones, instruments that monitor the oceans at all depths, other fixed and mobile, ubiquitous measurement systems will need to be devised to monitor all of this. Critically, data collection must be both guided and interpreted by theory and analytics. This involves communication systems, data storage, computational systems, and multiscale theory to design collection processes, interpret, predict and intervene.

VI. Human-Biotechnology Interface

There are many aspects of a program to advance biotechnology that touch on sociological issues. First, the NSF MPS should pay attention to aspects of the research community that impact its ability to make progress along the lines outlined above. At a deeper level, we should also pay attention to perceived societal impacts of the technology and engage with the corresponding communities to ensure sensitivity to the opinions of the public at large.

6.1 Collaboration: The MPS division and, more broadly, the NSF needs to develop additional support mechanisms that promote collaborative research involving diverse scientists with the expertise and tools required to address big problems. Whereas interdisciplinary research involving productive collaborations between two or three principal investigators from different disciplines are becoming widespread, there remains a need for additional funding mechanisms for collaborations involving larger groups of scholars, particularly at the MPS-biotech interface. Such collaborations are becoming increasingly essential to tackle big problems, most of which require diverse strands of exploratory research, a subset of which warrant consolidation in a data-driven manner into the most promising potential solutions to the problem. An MPS-LW presentation by Twist Bioscience highlighted the importance of involving the private sector in such collaborations.

6.2 More interdisciplinary research and training centers in biotechnology: The inherently interdisciplinary nature of biotechnology lends itself well to making it a venue to foster cross-disciplinary collaborations as well as training. New Centers are needed, akin to MRSECs, which have been catalyzed the field of materials science and engineering, and Centers for Chemical innovation (CCI), which have done the same for chemistry. To further attract talent, the highly successful NSF Graduate Research Fellowship Program (GRFP) should be expanded to specifically train a cadre of graduate students with one foot in the life or medical sciences and another in the physical sciences. These NSF-funded biotechnology research centers should be given wide latitude to pursue high-risk research. The fast-paced nature of biotechnology requires real-time identification and nurturing of nascent breakthroughs that have potential to achieve paradigmatic change. These timescales are incompatible with traditional mechanisms for grant preparation, application, revision, funding, and initiation.

6.3 Sustainability, Environmental Impact, and Ethics: Finally, there should be investment in research focused on sustainability, environmental impact, and ethics. Questions that focus on addressing challenges in developing sustainable manufacturing practices that benefit biotechnology are needed. As one example, ethics connects to biotechnology in ways

that determine possible uses of CRISPR technology to ameliorate disease by engineering our own germline genomes. To the extent the NSF invests in interdisciplinary biotech centers, a portion of each center's budget should be reserved for activities of this nature.

6.4 Data Sharing Issues: Societal boundaries have been erected to protect tool developers, allowing them some level of time-limited exclusivity for their creations. While limited exclusivity does provide a needed advantage in order to spur innovation, it likewise creates circumstances where duplication of effort becomes a constant occurrence.

Focusing on tool development and taking a panoptic view of this yields the following insights:

- having an active up-to-date past roadmap of prior tools during the 20-year investment period is necessary to minimize duplication of effort
- both successful and inconclusive results of projects should be maintained, by providing a way to publish inconclusive results as part of the public record, thereby supporting the prevention of duplication
- leveraging machine learning to show cross-connects between inputs and outcomes of projects, then following suggested development areas identified by machine learning for projects in unexplored or underexplored areas of scientific knowledge
- training and using proposal reviewers who favor higher-risk, higher-outcome projects, where such reviewers may not even recognize their risk-averse behavior and have unconscious bias against risk
- use fuels success, therefore, industry adoption should be accelerated through matchmaking between developers and industry through machine learning based on the past roadmap and the results.

The NSF currently succeeds by promoting fundamental scientific and engineering investigations in many disparate fields. It hosts a compendium of knowledge, where all outcomes of scientific projects funded by NSF can be set into a public record, whether successful or inconclusive. This knowledge base provides both metrics and large-scale data for machine learning to elucidate key predictors highlighted above over the 20-year period under consideration. It seems tantamount to success to know as much about what does not work and what has already been accomplished when considering subsequent project proposals.

VII. MPS and the Living World Subcommittee

Rommie Amaro, University of California, San Diego

Virginia Cornish, Columbia University

Sara Del Valle, Los Alamos National Laboratory

Nebojsa Duric, University of Rochester

Moh El-Naggar, University of Southern California

Anthony Guiseppi-Elie, Anderson University

Trachette Jackson, University of Michigan, Ann Arbor

Chaitan Khosla, Stanford University, **Co-Chair**

Ka Yee Lee, University of Chicago, **Chair**

Herbert Levine, Northeastern University, MPS AC Member

Jennifer Lewis, Harvard University, MPS AC Liaison to Subcommittee

Michael Murrell, Yale University

Sherine Obare, North Carolina Agri. & Tech. State Univ., Univ. of North Carolina, Greensboro

Jeff Pixton, National Radio Astronomy Observatory

Dagmar Ringe, Brandeis University

William Tolman, Washington University, St. Louis, MPS AC Member

Eranthie Weerapana, Boston College

Neal Woodbury, Arizona State University, **Co-Chair**

Linda Sapochak, National Science Foundation, MPS Liaison to Subcommittee

Leighann Martin, NSF Office of MPS Assistant Director

VIII. Appendix - Workshop Details



Future of Biotechnology Workshop June 29 - June 30, 2021

Day 1 – June 29, 2021, 1:00-4:15pm EDT

Zoom Link: <https://nsf.zoomgov.com/j/1605293424?pwd=bHRic3lDWE8vcGY5ZWpMMnN2SUEzQT09>

Meeting ID: 160 529 3424

Passcode: 994131

1:00-2:45pm Big Picture Talks and Panel

Guest Speakers:

- [Holden Thorp](#) (*Science*)
- [George Church](#) (Harvard Medical School)
- [Abraham Stroock](#) (Cornell University)
- [Robert Nelsen](#) (ARCH Venture Partners)

1:00-2:00pm Short Talks

2:00-3:00pm Panel Discussion

3:00-3:15pm *Break*

3:15-4:15pm Closed Session for Discussion (Subcommittee Members Only)

Day 2 – June 30, 2021, 2:00-6:00pm EDT

Zoom Link: <https://nsf.zoomgov.com/j/1617845501?pwd=dmFtZkNFVkJZBRDdjVHlrVEJOK1BIUT09>

Meeting ID: 161 784 5501

Passcode: 898069

2:00-3:00pm: Subgroup Topic – Tools for Biotechnology: Theory and Computation

Guest Speakers:

- [Zaida \(Zan\) Luthey-Schulten](#) (University of Illinois at Urbana-Champaign)
- [José Onuchic](#) (Rice University)
- [Mary Jo Ondrechen](#) (Northeastern University)

3:00-4:00pm: Subgroup Topic – Critical Applications and Problems in Biotechnology

Guest Speakers:

- [Kelly Stevens](#) (University of Washington)
- [David Mooney](#) (Harvard University)
- [James J. Collins](#) (Massachusetts Institute of Technology)
- [David R. Walt](#) (Harvard Medical School; Brigham and Women's Hospital)

4:00-4:15pm *Break*



Future of Biotechnology Workshop
June 29 - June 30, 2021

4:15-5:00pm: Subgroup Topic – Tools for Biotechnology: Imaging

Guest Speakers:

- [Mark Anastasio](#) (University of Illinois at Urbana-Champaign)
- [Joshua Vaughan](#) (University of Washington)

5:00-5:30pm: Subgroup Topic – Human-Biotechnology Interface

Guest Speaker:

- [Catherine J. Murphy](#) (University of Illinois at Urbana-Champaign)

5:30-6:00pm Closed Session for Discussion (Subcommittee Members Only)

*This workshop is organized by the **MPS and the Living World** subcommittee of the [NSF Directorate for Mathematics and Physical Sciences \(MPS\) Advisory Committee \(MPSAC\)](#)*

The aim of this workshop is to address how best to enable and transform the future of biotechnology, addressing the following questions:

- What are the fundamental science questions that need to be solved to advance the field?
 - What new tools and/or techniques are needed?
 - Where can convergence/cooperation advance the field?
 - What partnership can advance the field?
 - What do we need to train the next generation of scientists and innovators for the field?
-