

Reference ID: 11226903973\_Read de Alaniz

---

**Reference ID:** 11226903973\_Read de Alaniz

**Submission Date and Time:** 12/16/2019 3:47:12 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

*Consent Statement:* "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

#### **Author Names & Affiliations**

Submitting author: Javier Read de Alaniz - University of California Santa Barbara

**Additional authors:** Prof. Heather Maynard, UCLA; Prof. Craig Hawker, UCSB; Prof. Yi Tang, UCLA; Prof. Scott Shell, UCSB; Prof. Glenn Fredrickson, UCSB; Prof. Ambuj Singh, UCSB; Dr. Tal Margalith, UCSB; Dr. Paul Weakliem, Director CSC, UCSB.

**Contact Email Address** (for NSF use only): (hidden)

#### **Research domain(s), discipline(s)/sub-discipline(s)**

Chemistry and Biochemistry, Materials, Computer Science, Chemical Engineering, Biomolecular Engineering

#### **Title of Response**

Cyberinfrastructure Needs and Opportunities for the NSF Materials Innovation Platform and the proposed Biomaterials Innovation Collaborative

#### **Abstract**

The Biomaterials Innovation Collaborative (BIC) seeks to create a living community at the nexus of synthetic biology and materials focused on the rapid discovery of new biomaterials. This Materials Innovation Platform will serve as a Knowledge Hub for biological pathways and chemical synthesis strategies, supporting a flexible research workflow within an immense design space, and create a critical opportunity for developing large-scale cyberinfrastructure for data collection, curation, analysis via ML, storage, and high-speed access, which will scale through partnership with National biosynthesis-focused centers. Additionally, new tools will catalyze an intimate connection between experimentalists in disparate fields (synthetic biology, chemistry, and materials), theoreticians and characterization specialists, and computation and data science specialists while closing the gap between academic research and industry through the training of a new student workforce versed in industrial-scale high-throughput experimental design and data management and evaluation.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

The Biomaterials Innovation Collaborative (BIC) seeks to create a living community at the nexus of synthetic biology and materials focused on the rapid discovery of new biomaterials. This Materials Innovation Platform will combine high-throughput, automated bio- and chemical synthesis infrastructure with advanced materials characterization and multi-scale computation techniques to create a flexible workflow in which users can both design new materials from novel bioderived building blocks and work backwards from desired material properties to select the appropriate synthetic pathways. In addition to providing a library of starting materials and an extensive array of equipment to support the experimental work, the BIC will serve as a Knowledge Hub for biological pathways and chemical synthesis strategies by creating a database that combines characterization data (traces and imagery), simulation outputs, and workflow information captured through a laboratory information management system (LIMS) that will collect data from both the automated equipment and from electronic lab notebooks maintained by the user. Machine Learning algorithms will be applied to the metadata layer that connects these databases to facilitate the design cycle. The design space associated with this endeavor is immense. As an example of the scale involved, consider that the simple exercise of combining two monomers into an 80-mer length chain, in which both sequence and sub chain length impact the final materials properties, affords 1023 possible combinations. Our vision for the BIC is that researchers will access hundreds, if not thousands, of potential building blocks while targeting equivalent numbers of desired endpoints. Beyond the research and knowledge developed and stored at the BIC, the Collaborative will also interface with existing programs in the field, such as DARPA Living Foundries, in order to extend the development pipeline, further expanding on the critical “molecule to biomaterial” aspect and increasing the size and impact of the BIC’s materials and pathways libraries and workflow databases. The associated cyberinfrastructure resources must be able to accommodate the storage of this large quantity of data, connect multiple center-scale knowledge hubs, support a range of

metadata, and support rapid multi-scale simulations and advanced machine learning applied to many distinct properties from the molecular to macroscopic scale, all while being secure and accessible by a large user base with a broad range of experiences and geographical locations.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).** Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

Computation for Simulation and Machine Learning: For both molecular simulations and field-theoretic simulations that serve as the cornerstones of the BIC's hierarchical computation platform, GPUs have become an indispensable tool that improves performance by 30-40x speedup. This dramatic increase in computational speed is transformative, not simply enabling faster calculations, but opening opportunities to ask new kinds of research questions – for example, moving from investigations of single system behavior to systematic exploration and automated design of systems using iterative simulations guided by machine learning and optimization algorithms. To support the Collaborative, we plan to add nearly a dozen GPU nodes to UCSB's Center for Scientific Computing. While campus-level shared computing resources are critical to the research community, the costs of maintenance, logistics, security, personnel, and environmental factors are expected to increase as the BIC scales through partnerships with other biosynthesis-focused centers. Expansion beyond local resources to national computing centers, large scale data repositories, and commercial clouds will be needed, along with funding models that are both sustainable and mutually appreciated by both NSF and campus, especially in relation to the differences between capital expense versus ongoing operational expense (including staff management and training resources/personnel for users) and between supporting local infrastructure versus partnering with cloud-based service providers.

Storage and Data Management: The quantity of technical data and vast array of data modalities (images, workflow metadata, simulation and characterization outcomes) will require a significant investment in storage infrastructure, setup, management, and upkeep/archiving, in order for the data to be available for mining and ML. Local storage can accommodate approximately a petabyte of data; larger archives will have to be stored in the cloud, which may limit access speed required for real time ML-based experimental design. It would be beneficial to have dynamic storage solutions that expand as required, are managed in a dedicated and cost-effective manner, and are accessible on-demand. Beyond storage and retrieval considerations, data curation and qualification also pose a critical national challenge. To address this in the BIC, we will engage the user base in the curation process by requiring users to detail how they will evaluate, curate, and score the quality of their data, and following up with a method that allows the user score to be

machine learned, allowing for user scores to be standardized. We will also task a Data Committee to oversee data management and curation. However, better guidelines and best practices are required to avoid stifling the data flow in favor of rigorous checkpoints for data quality. Although this approach will be implemented in the proposed BIC, the ability to scale best-practices for data curation, qualification, search, and retrieval across U.S. research agencies is needed to establish standards that can connect multiple center-scale knowledge hubs. **Networking:** There have been a number of initiatives to enable and improve high-speed connections for data transfer, including The California Research and Education Network which enables a 100Gbps connection between the BIC's nodes at UCSB and UCLA and between the UC system and other academic systems across California, and the NSF-supported Pacific Research Platform connecting our campuses to 50 national and international partners. While the BIC's mission will initially be based on in-person research and service requests, we envision a future in which users outside of UCSB and UCLA will engage in real-time collaboration with the BIC's resources including accessing and operating equipment remotely, further increasing demands on network capacity. However, the 'last mile' to the campus and within research buildings between equipment, storage, and computing resources are often bandwidth bottlenecks; this is an aspect of CI that should be considered critical for any future project.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

**CI Democratization:** To enable ML-assisted forward and inverse materials discovery and design, the BIC will be leveraging data collection and organization tools that are common in industrial settings but non-commonplace in academia, namely the LIMS system and electronic notebooks. There is a critical need in making these types of tools highly flexible and able to accommodate the wide range of workflows and data that will be produced in academia. An integral part of training a student workforce versed in industrial R&D and manufacturing practices will involve developing an expertise in the use of these tools and in the analysis of the large data sets produced in the course of their use. Additionally, a successful implementation of the Design-Build-Test-Learn (DBLT) cycle requires an intimate connection between experimentalists in disparate fields (synthetic biology, chemistry, and materials), theoreticians and characterization specialists, and computation and data science specialists. We must continue to build on existing efforts in training our researchers in database and metadata setup, ML and big-data analysis, integration of computation and experimentation in the DBLT cycle, and data curation for database integrity. This will require the development of undergraduate and graduate curricula, and the continuous engagement of experts at the interface of scientific domains and ML in training researchers.

-- End Submission --