Response to NSF 20-015, Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

Reference ID: 11222726180_Zhang

Reference ID: 11222726180_Zhang
Submission Date and Time: 12/14/2019 2:04:41 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: Peihong Zhang - Physics Department, University at Buffalo, SUNY
Additional authors: None

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

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

Physics; condensed matter physics; materials science

Title of Response

On materials prediction and design using machine learning methods and the development of standardized, easy-to-access, and accurate databases

Abstract

The last few years have witnessed a surge in exploring the possibility of using machine learning for understanding the properties and behavior of materials, and even for materials prediction and design. Most current research in this area use almost exclusively locally and specially generated data by the
research group to train the models. With many materials related open-access databases available, it is rather disappointing that there has not been much effort on exploiting existing database for materials discovery.

**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.

Exploiting the power of modern computers to accomplish complex tasks without specific instructions has become a reality in some areas with the help of various machine learning approaches. Making all this possible are the availabilities of (accurate) sample data and well-conceived mathematical models. Machine learning has increasingly been recognized as a powerful concept in materials design and predictions. However, there are many obstacles that need to be overcome before we can see meaningful achievements. It is imperative that we develop standard, easy-to-access, and reliable (accurate) database for the research community.

**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?

Unfortunately, except for a few rare cases, materials complexity almost guarantees the non-existence of a universal so-called structure-function relation. Practical applications of machine learning for materials prediction has been increasingly narrowed down to very specific materials systems with well-defined properties that are often represented by a single data entry (value). The model and understanding thus developed have nearly no value to other situations with even very minor changes to the assumptions and constraints. With many materials related open-access databases available, it is rather disappointing that there has not been much effort on exploiting existing database for materials discovery. I think there is some kind of disconnection between the database developers and the users (materials researchers). Developing standardized, easy-to-access, and accurate database is imperative. It is unlikely that a single database will fit all purposes. For example, the databases for research community focusing on optoelectronics will be very different from those for structural applications of materials. We also need
to develop meaningful models and descriptors that can truly represent the underlying materials and, at the same time, can be related to the properties of the materials.

**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.

-- End Submission --