

Reference ID: 11227312576\_Zunger

---

**Reference ID:** 11227312576\_Zunger

**Submission Date and Time:** 12/16/2019 6:57:04 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: Alex Zunger - Colorado University at Boulder

**Additional authors:** Zhi Wang, Colorado University at Boulder; Oleksandr Malyi, Colorado University at Boulder; Carlos Mera Acosta, Colorado University at Boulder

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

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

Materials Science; Physics;

#### **Title of Response**

Data-focused Cyberinfrastructure - response from Alex Zunger's group

#### **Abstract**

Alex Zunger has a NSF-DMR-DMREF project that has been previously approved and is being funded as: Project name: DMREF: Collaborative Research: Complex Nanofeatures in Crystals: Theory and Experiment meet in the Cloud. Reviewed and accepted by NSF-DMR, project date October 1, 2019 –

September 30, 2023, award number: 1921949 In this survey, we provide typical details on the improvements which are needed from the computational side for our NSF funded project. Specifically, we point out the importance of long-term data storage, a queuing system for data generation for machine learning and big data analysis, and access to high memory nodes. We also point out how the current NSF funded XSEDE resources can be improved further for our project specifically. The implementation of the suggestions discussed in this survey will allow us to implement physical models based on machine learning for the prediction of materials properties and data dissemination in general.

**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.

In recent decades, the accelerated computational production of data for individual systems has promoted the discovery and understanding of physical properties. The impact of these data-intensive “high-throughput” methods, which rely on massively parallel computing, large data storage and fast data I/O, is plausible in many areas of physics. For example, in condensed matter and materials science, there are remarkable examples including the computational prediction of all potential topological materials known in nature, and all two-dimensional materials that could be synthesized. Despite the success of “high-throughput” method in predicting single properties, this trial-and-error process is undoubtedly tedious and expensive (much worse in the prediction of multiple functionalities). Meanwhile, the rational search and prediction of functionalities and co-functionalities based on design principles (“inverse design”) require much less calculation. However, such “inverse design” requires a deep understanding of the physical mechanism determining the existence and magnitude of the desired property. “Machine learning” (ML) emerged as a solution from computer science for prediction of relationships between a specific property of a finite group of systems and their attributes (“descriptors”). Although the descriptors are intended to allow accelerated prediction of properties, this learning process is still data-intensive and always limited to the number of available systems and the used attributes. Two mutually related problems (challenges) arise in the use of ML techniques to the prediction of physical properties: i) transferability and ii) interpretability of the descriptor. Specifically, the ML model trained with a group of materials may not describe the physical property in a larger or inhomogeneous group of compounds. Additionally, the existence of a correlation between the descriptor and the physical property does not denote causality, in the most desirable scenario; the descriptor should allow the understanding of the physical mechanisms behind the property and its rational prediction. One of the possibilities to address this problem is to integrate the concept of inverse design with ML techniques, i.e., the introduction of design principles (or empirical physical model) as attributes in the learning process, rather than the commonly used properties. Thus, we consider that in cross-disciplinary solutions combining ML with e.g. physics, one of the greatest challenges is the development and implementation of physical-model-based ML for the rational prediction of

functionalities and the establishment of physical relations, which are previously hidden in traditional analysis and ML techniques.

**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?

From a computational perspective, developing big data analysis and machine learning requires improving in two main directions: Computing: Since the main calculations are performed in a high-throughput manner, it is important to develop different queuing systems, allowing doing fast calculations with a small queuing time. From our experience with NSF funded XSEDE resources, the queuing option is rather limited, which often increases the time for the “short” calculations. The high-throughput studies require the number of calculations to be performed at the same time and a high limit for the number of calculations that can be run/submitted at the same time. From our experience, for NSF funded XSEDE resources, the number of calculations which can be submitted from the same user is rather limited (e.g., Stampede2 only allows submission of 50 jobs per user account). For the specific cases, the calculations often do not require the high number of nodes, but often require high memory nodes. Because of this, it is important to have an option for running high memory nodes for specific calculations. Data storage: For developing machine learning and big data analysis, it is highly important to have large permanent storage, where the data can be stored for a number of years, as it allows to form the data bank which can be used for dissemination of the results as well as machine learning models. High-speed writing/reading for the data is highly important as machine learning often requires reading the data from hundred thoughts of calculations at the same time as well as generation of all these data with first-principles calculations. However, we also suggest that it is necessary to have a dynamic balance of I/O bandwidth for all users, so that the heavy I/O from one (or few) user(s) will have negligible effect on other users. General: We believe that it will be beneficial for fundamental solid-state physics/chemistry to get some basic allocation for each NSF funded projects (e.g., 2 million core-hours per year) on the XSEDE proposal without the application for the separate computer proposals.

**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.

We would suggest that a standard data “bank” for important materials that contains verified, standard input and output files of modern material simulation software (e.g. VASP and Quantum Espresso), will benefit the physical, chemical, and engineering communities. We already showed in our publications [Matter 1, 33-35 (2019); Materials Horizons, 6, 2124-2134 (2019); Nature 566, 447-449 (2019); Materials Today (2019) DOI: 10.1016/j.mattod.2019.08.003] that for many materials, e.g., the famous Mott insulators, the novel topological insulators, and the very promising photovoltaic materials, there are many theoretical studies using naïve models without giving details who gave results violated to experimental observations (e.g. metallic in theory but insulating in the experiment) and led to incorrect conclusions. Without such a standard data “bank” that collects the verified, “correct” theoretical results, researchers can easily get trapped by those fake results and miss the right conclusions. It is hence a heavy waste of data and computing resources of NSF. Such standard data “bank” is significantly different from existing open-access databases (e.g., Materials Project, Open Quantum Materials Database, and AFLOW database) as it aims to provide: (1) Fully open access data for verified theoretical studies; (2) full links between computed results and those reported in different literature; (3) full links among different open-access databases which are changing continuously and are usually uncorrelated to each other; (4) opportunity for different teams to provide the full data on their published results as well as to monitor existing literature; (5) guiding for future research to develop a simple pathway in the maze of available data; (6) economy of NSF resources in general.

-- End Submission --