NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 18-075

Dear Colleague Letter: Data-Driven Discovery Science in Chemistry (D3SC)

May 1, 2018

Dear Colleagues:

In 2016, NSF unveiled a set of "Big Ideas" — 10 bold, long-term research ideas that identify areas for future investment at the frontiers of science and engineering. Among them, the *Harnessing the Data Revolution* idea aims to promote engagement of NSF's research community in the pursuit of fundamental research in data science and engineering, the development of a cohesive, federated, national-scale approach to research data infrastructure, and the development of a 21st-century data-capable workforce[1]. Within this context, the Division of Chemistry (CHE) launched the *Data-Driven Discovery Science in Chemistry (D3SC) initiative in 2017 to support research* activities that seek to capitalize on the data revolution and promote data-driven discoveries to advance fundamental understanding of complex chemical systems[2].

The amount and variety of data generated in the chemical sciences, and the rate at which it is being produced, are rapidly increasing, driving the need for corresponding growth in our ability to extract useful insight from interrelated sources. In order to meet this challenge, chemical researchers must effectively share, mine, and repurpose their rapidly-growing chemical datasets in order to apply state-of-the-art data analysis tools to expand chemical understanding and make modern data science an integral part of the chemical research of the future. The report from a recent NSF CHE workshop on *Framing the Role of Big Data and Modern Data Science in Chemistry* identified many challenges and opportunities in this area[3].

The Division of Chemistry, together with the Catalysis Program and the Process Systems, Reaction Engineering, and Molecular Thermodynamics Program of the Division of Chemical, Bioengineering, Environmental, and Transport Systems (CBET)[4] through this Dear Colleague Letter (DCL) invite research proposals that utilize modern data science in the context of chemical and chemical engineering research. Successful D3SC proposals will emphasize

new information that can be obtained from better utilization of data (including data from multiple laboratories, techniques, and/or chemical systems), and how this can lead to new research directions. Proposals that foster and strengthen interactions among chemists and data scientists, and that jointly engage theory, modeling, and experimentation to advance research goals are strongly encouraged. The most competitive proposals will provide detailed discussion of specific data-enabled approaches to be used, the significant chemical problem to be studied, new fundamental chemical knowledge to be gained and the broader relevance of the proposed activities to other areas of chemical research. Proposal elements that consider error and uncertainty analysis, record and store appropriate metadata, and determine the robustness and reliability of data are encouraged. Examples of possible topics include (but are not limited to) using tools of data visualization, data mining, machine learning (including emerging approaches such as deep learning and active learning), or other data analysis approaches to:

- Accelerate the discovery of homogeneous or heterogeneous catalysts with improved activity and selectivity, as well as the discovery of new catalytic transformations;
- Advance the design of new chemical species and/or synthetic reactions, and forecast improved synthetic conditions;
- Map the mechanisms by which chemicals interact and transform, both covalently and noncovalently, and predict structure/property relations based on existing chemical datasets;
- Discover principles of multiscale organization underlying emergent chemical phenomena in macromolecular systems;
- Enable real-time feedback loops between chemical data collection and processing for rapid identification and correlation of key events during chemical measurements;
- Harness chemistry's rich, diverse but distributed datasets and identify novel ways of sharing and utilizing chemical data derived from multiple instruments, datatypes, and locations;
- Develop innovative approaches for integrating, correlating, and analyzing chemical simulation or measurement data to provide new chemical insights.

Note that the construction or maintenance of large-scale databases *per se* is not the focus of this DCL, although such databases may be required as a means to the endpoint of using the data to provide insights and predictions. Proposals focused on developing cheminformatics for biomedical or materials research applications are outside the scope of this DCL. Proposals whose primary focus is on the development of algorithms and software should be submitted to the Computational and Data-Enabled Science and Engineering (CDS&E) program[5]. Proposals on the development of general-purpose data mining or analysis algorithms not aimed at addressing a specific chemical question are more appropriate for programs supporting general tool development outside of the CHE division. Researchers seeking support to build data infrastructure and establish long-term data capabilities are

encouraged to consult with the Cyberinfrastructure for Sustained Scientific Innovation (CSSI) program on Data and Software[6]. Team research proposals seeking to identify potential new research areas that go beyond what is described in this D3SC DCL are encouraged to consult the Dear Colleague Letter: Growing Convergence Research[7].

Proposals in response to this DCL should be submitted to the existing program of interest in CHE[8] or selected programs in CBET[4] during the regular submission windows (deadlines) of the corresponding programs. The proposal title must be tagged with "D3SC:". Other than the proposal title, the NSF Cover Sheet should be prepared as a regular proposal submission to the program. Principal Investigators (PIs) are strongly encouraged to contact the cognizant D3SC Program Officers[9] prior to submission to determine the appropriateness of the work for consideration.

Proposals may be submitted in combination with other solicitations. For example, proposals may be submitted in combination with the Faculty Early Career Development Program (CAREER)[10] or with the Facilitating Research at Primarily Undergraduate Institutions: Research in Undergraduate Institutions (RUI) and Research Opportunity Awards (ROA) solicitation[11].

If there are strong collaborations with industry, the Grant Opportunities for Academic Liaison with Industry (GOALI)[12] proposal type can be used in conjunction with this effort. These proposals should be submitted to the appropriate solicitation with "D3SC" added to the title (for example, CAREER:D3SC:, RUI:D3SC:, or GOALI:D3SC: Name of your proposal). Proposals including international collaboration are encouraged when those efforts enhance the merit of the proposed work. NSF typically supports the costs of the U.S. team and foreign partners are typically supported by their own funding agencies.

Submission of other types of proposals such as EAGER[13] (EArly-concept Grants for Exploratory Research) and RAISE[14] (Research Advanced by Interdisciplinary Science and Engineering) proposals may also be appropriate, but principal investigators are required to contact one of the cognizant D3SC Program Officers[9] for additional guidance in advance of a potential submission. For EAGER or RAISE proposals, the title of the proposal should have "EAGER:" or "RAISE:" specified, followed by the "D3SC:" designation. Requests for supplemental funding may also be appropriate; again, please check with the cognizant D3SC Program Officers[9] for additional guidance. EAGER, RAISE, and supplemental funding requests can be submitted at any time but are encouraged by **April 15, 2019, 5:00 pm**, submitter's local time, to ensure timely consideration.

All programs in the Division of Chemistry and the Catalysis Program and the Process Systems, Reaction Engineering, and Molecular Thermodynamics Program of CBET[4] are

excited by the opportunities in the D3SC area and look forward to working with the community to develop new approaches to gain insights from existing data, as well as new experimental and theoretical results. For recent D3SC awards, please search the NSF award database with the keyword "D3SC"[2]. For general questions about this DCL, email the cognizant D3SC Program Officers[9].

Anne L. Kinney
Assistant Director
Directorate for Mathematical and Physical Sciences
National Science Foundation

Dawn M. Tilbury
Assistant Director
Directorate for Engineering
National Science Foundation

References

1. Harnessing Data for 21st Century Science and Engineering" in 10 Big Ideas for Future NSF Investments.
2. CHE D3SC Awards.
3. CHE Data Workshop Report.
4. The Catalysis program and the molecular thermodynamics aspects of the Process Systems, Reaction Engineering, and Molecular Thermodynamics (PRM) program within CBET: https://www.nsf.gov/funding/programs.jsp?org=CBET.
5. Computational and Data-Enabled Science and Engineering (CDS&E).
6. Cyberinfrastructure for Sustained Scientific Innovation (CSSI) - Data and Software.
7. NSF 18-058, Dear Colleague Letter: Growing Convergence Research.
8. CHE research programs.
9. D3SC cognizant Program Officers: for CHE: Lin He (lhe@nsf.gov), Ken Moloy (kmoloy@nsf.gov); for CBET: Robert McCabe (rmccabe@nsf.gov), Triantafillos Mountziaris (tmountzi@nsf.gov).
10. NSF Faculty Early Career Development Program (CAREER).
11. Facilitating Research at Primarily Undergraduate Institutions.
12. Grant Opportunities for Academic Liaison with Industry (GOALI).
13. EArly-concept Grants for Exploratory Research (EAGER).
14. Research Advanced by Interdisciplinary Science and Engineering (RAISE).