Building MSU towards global recognition for excellence in research and training in biological data science and computational biomedicine

***Strategic Plan Theme*: Global Impact**

***Funding Level*: More than $5 million**

***Facility Needs*: New facilities will be needed**

***Submitting Unit*: CMSE**

Collaborating colleges/departments/units involved with this proposal.

Dept. Computational Mathematics, Science and Engineering, Dept. Biochemistry and Molecular Biology, Dept. of Pathobiology and Diagnostic Investigation, and Dept. of Microbiology and Molecular Genetics.

What is the proposal's big theme or idea?

Big data, artificial intelligence, and computing are revolutionizing biology, agriculture, and medicine. The idea behind this proposal is to put MSU at the forefront of this revolution.

What is the proposal's goal?

This proposal's goal is to establish an integrated infrastructure and environment that sustains a virtuous cycle between: I. Attracting and retaining high-caliber students and researchers in the areas of bioinformatics, computational biology, and biomedical data science, and II. Sustaining an excellent training and research program that drives: - technological innovation in data science, machine learning, and artificial intelligence, - knowledge discovery at all levels of biological organization, and - translation to applications in agriculture and healthcare.

A roadmap for achieving this goal and becoming one of the national and global leaders in the areas of biological data science and computational biomedicine includes the following 5 milestones: 1. Training existing researchers to analyze high-throughput data. 2. Offering bioinformatics and computational biology electives to undergrads and graduate students. 3. Setting up a formal MS/PhD program in biological data science and computational biomedicine. 4. Building up a biological data analysis service unit to enable a connected compbio community. 5. Building a center of excellence involving multiple faculty hires and the creation of a new center.

Define the significance, or impact of your big idea.

Biology and medicine are being fundamentally shaped by data, algorithms, and computing. With rapidly decreasing costs of experimental technologies, biological data is being generated at an exponential rate and set to overtake any other type of data (including astronomical data) over the next decade. Medicine too has a data-rich future with the generation of a massive data library — genetic sequences, molecular profiles, data from wearable devices, images, and electronic health records — for every individual walking into the clinic. Effectively leveraging these biological and medical data requires: i) the development and implementation of new machine learning and artificial intelligence technologies and tools, and ii) the comprehensive application of these techniques and tools to specific biomedical problems and translation to novel diagnosis and treatment opportunities.

At this moment, MSU has the potential and the opportunity to make a significant contribution to this future and gain global recognition in this area by building on the effort over the past 7 years to recruit experts in biological data science and computational biomedicine through several initiatives: - Hiring multiple tenure-track faculty in these areas (esp. via the Global Impacts Initiative), - Establishing the Department of Computational Math, Science and Engineering (CMSE) and the Institute for Quantitative Health Science and Engineering (IQ), - Ensuring the presence of computational faculty at Grand Rapids (facilitating their partnerships with Van Andel and Spectrum Health). A concerted campus-wide initiative is also crucial to support the two new partnerships — i) the Henry Ford Health + MSU Health Sciences Henry-Ford and ii) the McLaren-MSU Health Campus — that can open the doors for clinical research opportunities, previously not possible due to the lack of MSU-affiliated hospitals and health systems.

Who will be impacted?

MSU is the home of 110 NIH-funded projects from 15 different institutes — together amounting to >$47.6 million — that involve the generation or analysis of massive biological sequencing data to study a wide range of problems in biology and biomedicine including gene regulation, development, neuroscience, cancer, and drug development. Effectively analyzing these data fundamentally requires sophisticated computational analysis supported by MS- and PhD-level scientists who can work with the groups as needed. Currently, such a service is nearly completely lacking at MSU and it is stifling a number of research programs across the EL and GR campuses.

Furthermore, >25 labs at MSU are developing methods and tools in the areas of bioinformatics and computational biology along with growing demand in biomedical data science. However, we do not have one graduate program that provides holistic training in these areas. >25 relevant courses do exist, but they are fragmented across departments. A single program is also crucial for attracting potential students/postdocs looking for training and research in these areas and providing them with a home.

The ideas outlined in this proposal are also crucial for supporting the clinical research opportunities that are emerging thanks to the new partnerships with Henry Ford Health and the McLaren-MSU Health Campus.

This effort is crucial for putting MSU on the map in data-driven biology and medicine and making it competitive with other universities — including the University of Michigan — in advancing training and research in these areas.

What does sustainability for your proposal look like?

MSU has made great strides towards reaching Milestones 1 and 2, especially by establishing the Dept. of Computational Math, Science and Engineering (CMSE). Computational biology and bioinformatics faculty in CMSE (including AK) have developed a comprehensive curriculum of bioinformatics modules and semester-long computational biology courses that have now been taught to >350 trainees at MSU across several units/colleges. The IMPACTS program (via a 2018 NSF training grant) has made additional headway by focusing on training in computational plant science. R-Ladies (led by JR) has been instrumental in data science education in non-traditional settings.

The next step (Milestone 3) is to establish a graduate program in biological data science and computational biomedicine. Following program growth leveraging initial investment from MSU, the target would be to expand to postbac and postdoctoral programs and to apply for NIH/NSF training grants.

Milestone 4 is to build up a centralized, shared biological data science and computational service unit to serve a connected computational biology community. Such a service needs to meet a variety of needs, which in turn requires a few different staff members in this service team (and computational infrastructure): 1. Bioinformatics system administrator and computing consultant: Help researchers effectively use (ICER- or researcher-owned) dedicated bioinformatics servers, the HPCC, and cloud infrastructure by installing and updating software, managing (open and HIPAA-compliant) data storage and access, and helping to establish efficient analysis workflows. 2. Bioinformatics and computational biology staff scientists: Masters and PhD-level scientists who can work with individual researchers across campus in various ways. 3. Software developers: Lead the development of distributable code and software packages 4. Web developers: Lead the

development of production-ready interactive web-interfaces and web-servers for research data and algorithms using local and cloud computing services. Following the initial setup to leverage MSU's investment, most of the long-term support for this service unit should come from extramural grants (where researchers write in their staff member support).

The final goal (Milestone 5) is to integrate the faculty and trainees across campus engaging with the graduate program and the service unit (Milestones 3 & 4) into a cohesive center of excellence. It would be ideal to establish a physical center that brings together diverse colleagues from various units. Such a center will present numerous opportunities to build collaborative teams, establish clinical research partnerships, recruit new faculty at various levels, provide a holistic home for trainees in biological data science and computational biomedicine, and apply for center grants.