Technological advancements have enabled the generation of incredible amounts of massively parallel “omics” data that describe the presence and abundance of genes (genomics, transcriptomics), epigenetic modifications (epigenomics), proteins (proteomics), metabolites (metabolomics), and microbes (metagenomics) in diverse biological systems. Analysis of omics data presents an opportunity to take a true systems biology approach to answering research questions that has not been possible to this extent in the past. Specifically, omics data provides unique insight into the structure and dynamics of entire systems at multiple levels from individual cells to whole organisms and even ecosystems, and parallel analysis of these data serves to form the big picture of these systems. However, standard approaches for analyzing these data rely on inferential statistics, which are not sufficient to grasp the dimensionality of data and intricate network associations of different variables. Moreover, the methods of analysis used vary between analyses and data types, which may result in drawing false equivalencies.

My broad research interest is to further develop high-resolution methods for the analysis of biological big data into a standardized methodology and tool (i.e., platform) for use across effectively infinite biological systems. I plan to achieve this by expanding the machine learning pipeline developed through my graduate research that was ground-truthed with omics data generated from non-model organism and agriculturally-important fish species. In doing so, I aim to answer and help others answer research questions designed to inform policy and practice in agriculture- and biomedicine-related fields and present standardized methods of analyses to the greater research community.

The deliverables of my research thus far have set the foundation for achieving this; I have successfully developed an analysis pipeline beyond the testbed stages that incorporates data mining strategies, machine learning models, and cross-validation methods, and is not limited by the constraints of inferential statistics. This pipeline proved to be successful at overcoming critical bottlenecks such as the reduction of data dimensionality. For example, over 21 million expression values (loci) were reduced to twenty-nine key genes that are distinct between groups of individuals when sorted by agriculturally relevant traits (e.g., growth). These genes correlate to a similarly reduced number of metabolites from a parallel dataset, and confirmed reports from other studies. Thus, these results led to the identification of targets of interest for selective breeding and genome engineering efforts. These target genes are to be incorporated into the breeding programs for these fish and are anticipated to be considered genes of interests for other selection or modification efforts across aquaculture species.

My expansion of this pipeline as a tool will be applied to two major goals of biological research in the 21st century: predictive phenotyping in agriculture and comparative omics between model and non-model organisms for biomedical research. The capabilities of this pipeline to identify unique loci determinant of specific phenotypes important for agricultural yields are increasingly vital for designing reproduction and rearing programs in the face of a rapidly increasing global population, climate crises, and subsequent food security challenges. The development of this pipeline to perform accurate comparative omics studies between model and non-model organisms is of similarly vital importance to biomedicine, as the ability to rapidly perform assessments

of and outcome predictions for medical interventions (e.g., in precision medicine) is a major limitation in developing such treatments.

The outlined objectives of my research will complement those currently being undertaken by many other researchers and lend themselves to a number of collaborations. The continued development of omics resources by research groups can be incorporated into beta-testing of the pipeline and comparative research between model and non-model organisms. My experience with and continued work on agriculturally-relevant species support the overarching goal held by many institutions, to understand biological systems and apply this knowledge to improving practices across critical industries such as agriculture and human health sciences. I anticipate this work will garner support and interest from many government agencies and departments such as the United States Department of Agriculture (USDA), National Institutes of Health (NIH), and Department of Defense (DoD), as well as private and non-profit groups and organizations. I look forward to working with the support and collaborative effort from such groups towards our shared goal of securing health and prosperity to our nation and the greater global society. Further, and as mentorship is a critical component of long-term success, I am committed to completing this work with a specific focus on training the next generation of the workforce in biotechnology, bioinformatics, and applied science. All that I aim to do is ultimately only worth what it enables others to do.