LETTER FROM THE EDITORS: THE EMERGING NEED FOR BIOLOGICALLY INSPIRED AND MATHEMATICALLY GUIDED MACHINE LEARNING FOR KNOWLEDGE DISCOVERY IN BIOLOGY

Over the last decade, there has been a growing number of efforts to address the need for interdisciplinary research in mathematical biology and its critical role in fostering the realization of modeling, analysis, and simulations that facilitate the discovery of novel biological phenomena, rules, and theories. In this special issue, we reflect upon the importance of this call and the compelling reasons for promoting collaboration across disciplines through four diverse research articles.

The convergence of mathematics, biology, computer science, and machine learning (ML) has opened up unprecedented opportunities for scientific exploration. As datasets related to biological systems become increasingly complex, traditional analytical methods often fall short of capturing their intricacies. Integration of ML algorithms [including deep learning (DL)] and mathematical models has emerged as a powerful toolset, empowering researchers to analyze vast amounts of biological data, uncover hidden patterns, and generate novel insights. Furthermore, the connections between mathematical biology and ML facilitate the development of computational frameworks that can simulate, predict, and explore a variety of biological phenomena. By integrating ML into mathematical and computational models, researchers can enhance their accuracy, efficiency, and adaptability, leading to deeper insights and more reliable predictions.

This special issue is a part of the National Science Foundation Models for Uncovering Rules and Unexpected Phenomena in Biological Systems (MODULUS*) program that brought together a highly diverse group of individuals in August 2022, from applied mathematics, computer science, artificial intelligence, biotechnology, physics, and biology, including senior and junior faculty members, experts from government labs and funding agencies, as well as postdocs and graduate students. This diverse group helped to define emerging research areas in biological systems. One of the priority topic areas identified in the workshop was the need to design, develop, and deliver novel tools and techniques for biologically inspired and mathematically guided ML for next-generation knowledge discovery in biology. Under this topic area, the participants identified some challenges and needs, including why investing in this topic is urgent, and a collective summary follows.

- **Challenges and needs:** Biology today has become a data-intensive research field. With the massive surge in data volume (numbers of measurements) and dimensionality (numbers of separate things measured) in biology, understanding the foundational rules of life is becoming a reality. If the growth continues at the current rate by doubling every seven months, we should reach more than one exabase of sequence per year in the next five years and approach one zettabase of sequence per year by 2025. Proteomic, metabolomic, genetic, phenotypic, and imaging datasets are growing almost as quickly. Along with these,

*https://www.nsf.gov/pubs/2021/nsf21069/nsf21069.jsp

2689–3967/23/$35.00

datasets from infectious diseases need quick and better interpretation to make informed decisions and stop future pandemics such as COVID. This data explosion requires new tools involving ML/DL and mechanistic models with mathematical rigor for knowledge discovery. Currently, the interpretable and robust ML/DL models inspired by the datasets and questions from biology are lacking. Mathematically guided and more interpretable models are needed to integrate the increasing number of diverse types of biological datasets across different scales.

- **Why now:** Although ML/DL has produced powerful prediction tools for biology and many other domains, designing effective ML/DL models for specific biological problems requires expertise in biologically inspired design, computational implementation of neural networks, and rigorous mathematical analysis. The mathematical biology community is uniquely positioned to contribute to and fully utilize the advances of the ML/DL field. In many biological applications, satisfactory predictions can be achieved by existing ML/DL models. However, the potential of ML/DL models to generate new biological hypotheses is not entirely realized. Similarly, massive amounts of biological data at multiple scales and systems can enable innovative ML/DL research. On the other hand, the lack of "ground truth" from correctly labeled data impedes the validation of the models.

  To close the data gap, additional experimentation and data collection driven by close collaborations between biologists, mathematicians, and computer scientists are in urgent need. For example, depending on the scale of the datasets, one can develop either traditional ML or DL approaches to model the datasets for discovery. White-box mechanistic models and black-box ML/DL approaches may be integrated for more interpretable outcomes that help with hypothesis generation. Mathematical analysis of such new hybrid models will lead to a deeper understanding of the robustness of the tools and the datasets.

- **Summary:** The next generation of ML/DL biology research is data-driven, incorporating scientifically intuitive, biologically inspired, and mathematically validated learning. Studies are needed to evaluate the robustness of the developed models. There is a significant need for robust and interpretable ML/DL approaches to address specific problems in biological systems. The outcomes from this research direction can result in identifying biological mechanisms, discovering new biological principles, creating trained models with stable performance, and developing more efficient frameworks to solve biological questions.

This special issue consists of four diverse articles written by researchers across multiple disciplines who share their work at the intersection of interdisciplinary mathematical biology and ML. The articles represent the collective efforts of scientists striving to harness the potential of ML to revolutionize our understanding of biological systems.

One article employs a physics-informed neural network platform to investigate the post-pandemic impact of COVID to help predict the behavior of population subgroups that move between neighboring cities. Another article demonstrates the robustness of using physics-informed neural networks as a data-driven tool to calculate optimal vaccine distribution plans for informed policy decisions in heterogeneously mixed populations. The third paper uses a data-driven approach called NeuralGene, a neural ordinary differential equation-based model, to reconstruct continuous dynamic systems governing gene regulation from temporal gene expression data and show that the method can capture the temporal dynamics of gene expression and classify cell fate decisions. The fourth article focuses on gene co-expression estimation, a fundamental step

for downstream single-cell data analysis. The study presents the application of state-of-the-art gene co-expression estimation methods on two novel simulation processes, revealing the complexity of co-expression estimation for single-cell data and suggesting potential directions for improvement. Together, these four articles reflect a paradigm shift in interdisciplinary mathematical biology to promote the realization of modeling platforms that facilitate the discovery of novel biological phenomena, rules, and theories.

Guest Editors:

Ritambhara Singh
Department of Computer Science
Brown University
Center for Computational Molecular Biology
Brown University
USA

Padmanabhan Seshaiyer
Department of Mathematical Sciences
George Mason University
USA